2019

Rule Mining and Sequential Pattern Based Predictive Modeling with EMR Data

Orhan Abar
University of Kentucky, orhanabar@gmail.com
Digital Object Identifier: https://doi.org/10.13023/etd.2019.330

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Orhan Abar, Student

Dr. Ramakanth Kavuluru, Major Professor

Dr. Mirosław Truszczyński, Director of Graduate Studies

Rule Mining and Sequential Pattern Based Predictive Modeling with EMR Data

DISSERTATION

A dissertation submitted in partial
fulfillment of the requirements for
the degree of Doctor of Philosophy
in the College of Engineering at the
University of Kentucky

By
Orhan Abar
Lexington, Kentucky

Director: Dr. Ramakanth Kavuluru, Associate Professor of Biomedical Informatics
Co-Director: Dr. Jinze Liu, Associate Professor of Computer Science
Lexington, Kentucky 2019

ABSTRACT OF DISSERTATION

Rule Mining and Sequential Pattern Based Predictive Modeling with EMR Data

Electronic medical record (EMR) data is collected on a daily basis at hospitals and other healthcare facilities to track patients' health situations including conditions, treatments (medications, procedures), diagnostics (labs) and associated healthcare operations. Besides being useful for individual patient care and hospital operations (e.g., billing, triaging), EMRs can also be exploited for secondary data analyses to glean discriminative patterns that hold across patient cohorts for different phenotypes. These patterns in turn can yield high level insights into disease progression with interventional potential. In this dissertation, using a large scale realistic EMR dataset of over one million patients visiting University of Kentucky healthcare facilities, we explore data mining and machine learning methods for association rule (AR) mining and predictive modeling with mood and anxiety disorders as use-cases. Our first work involves analysis of existing quantitative measures of rule interestingness to assess how they align with a practicing psychiatrist's sense of novelty/surprise corresponding to ARs identified from EMRs. Our second effort involves mining causal ARs with depression and anxiety disorders as target conditions through matching methods accounting for computationally identified confounding attributes. Our final effort involves efficient implementation (via GPUs) and application of contrast pattern mining to predictive modeling for mental conditions using various representational methods and recurrent neural networks. Overall, we demonstrate the effectiveness of rule mining methods in secondary analyses of EMR data for identifying causal associations and building predictive models for diseases.

KEYWORDS: NLP, Machine Learning, Deep Learning, Association Rule Mining, Contrast Sequential Rule Mining, Causal Association

Author's signature: Orhan Abar

Date: July 31, 2019

Rule Mining and Sequential Pattern Based Predictive Modeling with EMR Data

By
Orhan Abar

Director of Dissertation: Ramakanth Kavuluru

Co-Director of Dissertation: Jinze Liu

Director of Graduate Studies: Mirosław Truszczyński

Date: July 31, 2019

To my parents Yeter and Mehmet, my wife Tugba and daughter Erva, and my brothers and sisters.

ACKNOWLEDGMENTS

While completing this dissertation, I have received great deal of support and assistance. First of all, I would like to express my deepest appreciation to my adviser Dr. Ramakanth Kavuluru for his invaluable contribution, suggestions, experience, and patience. I would like to extend my sincere thanks to my committee members: Dr. Richard Charnigo, Dr. Jinze Liu, and Dr. Dakshnamoorthy Manivannan for participating in my committee.

I gratefully acknowledge the help of Mehmet Bakal. I also had great pleasure of working with Dr. Gregory Heileman for my final year of study. I would like to thank the ministry of national education of Turkey for their full financial support. Finally, I would like to thank my family for all their relentless support during my studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

Increased digitization of data from various facets of our daily lives (including shopping runs, fitness activities, hospital visits, and social media interactions) necessitates new methodological advances in collecting, integrating, and mining extremely large datasets. Many data mining algorithms have been proposed during the past three decades in order to extract useful information from these large datasets. Such algorithms are currently used in fields such as chemistry, finance, e-commerce, biomedicine, and healthcare. In particular, the healthcare field has seen a major surge of applications of data mining mostly due to the deluge of digital data captured through electronic medical records (EMRs). However, this area poses significant challenges due to the high dimensionality (tens of thousands of variables), inherent errors, privacy concerns, and missing values. Our main goals are to leverage this EMR data to identify interesting associations between different biomedical variables of interest (potentially leading to new insights and hypotheses) and to build predictive models to identify high risk patients for chronic conditions. Next, we discuss the various patient attributes available in EMRs considered for this dissertation.

1.1 What is an EMR?

An EMR is a digital record that gets generated for each visit a patient makes to a healthcare facility (e.g., hospital, emergency room, diagnostic lab, private clinic). As such, each patient, as they go through a healthcare system, generates a trail (specifically, a temporally ordered sequence) of EMRs, one per visit. An EMR contains patient demographic information including the name of the patient, their gender, and age. Basic variables such as height and weight (hence body mass index (BMI)) and smoking status may also be recorded. More importantly, depending on the visit, an EMR may also have the following elements.

- At least one diagnosis code assigned from the standard terminology – *international classification of diseases: clinical-modification* standards (ICD-9/10-CM). These codes represent different conditions the patient has been diagnosed with during the visit.
- Depending on the nature of the visit, EMRs may also contain procedure codes from the *current procedure terminology* (CPT) standard for any procedures

performed (e.g., surgery) during the visit.

- If diagnostic lab tests (e.g., lipid panel) are done, values measured for different biomarkers (e.g., cholesterol) may be recorded typically using the *logical observation identifiers, names, and codes* (LOINC) terminology.

- Any medications administered during the visit or prescribed for subsequent usage will also be recorded via a standard terminology such as the *national drug code* (NDC).

- Besides these structured variables, free text is often included (especially for inpatients) in the form of admission notes, progress notes, pathology/radiology reports, and discharge summaries. Intuitively, the free text notes are expected to contain case-specific elaborate details that are not captured in any of the structured sources covered earlier (e.g., observed side affects, social variables including employment/marital status).

1.2 Applications of EMRs

As the digital encapsulations of a patient's health related events, EMRs are essential in improving the operational side of delivering patient care. EMRs are instrumental in maintaining continuity of care as patients go through different providers within a healthcare system. For instance, they can be crucial in ensuring patients obtain prescriptions that do not interfere with their existing medications. On the fiscal side, EMRs (esp. the structured codes) are also critical in determining what the patient or their insurance firm ought to be charged for the services the healthcare facility or the physician has provided.

Due to recent rapid adoption of EMRs among many facilities and better linking of records between different clinics that belong to larger systems, massive EMR datasets are being curated for millions of patients. If the healthcare system covers a reasonably sized neighborhood, one could argue that, the chronological aggregation of a patient's EMRs constitutes their longitudinal EMR (LEMR). Although there may be occasional visits to non-local clinics, given how insurance policies are tailored to minimize co-pay and other out-of-pocket expenses for in-network visits, patients are likely to limit most of their visits to a single healthcare system. The only exception to this assumption is when patients move permanently to a different location, which is easy to spot in their record based on the duration from their last visit to an in-network clinic. Thus, an LEMR is in essence an (ill)health trajectory of a patient's journey

through the healthcare system. Both visit-level EMRs and patient-level LEMRs are hence becoming goldmines for deriving insights across populations (as opposed to their main purpose: individual patient care). With the evolution of *data science* as a discipline and the rise of *precision medicine* initiatives for targeted therapies, (L)EMRs are being repurposed for disease phenotype discovery, predictive modeling, computational drug discovery and repositioning, cohort selection, and causal association mining. This secondary data analyses of EMR data for deriving insights at the patient population level is the main focus of this thesis. Next, we present a brief overview of this dissertation.

1.3 Overview of this Dissertation

This dissertation focuses on (1). assessing the potential of data mining approaches for extracting statistically significant and causal associations from EMRs; (2). predicting chronic conditions from LEMRs using recent advances in contrast pattern mining and deep neural networks. Although current approaches have been shown to obtain promising results, developing new approaches and/or cleverly configuring existing ones may create more predictive power and improve the quality of the outcomes as outlined in the following chronological introduction to the main ideas behind this dissertation.

- **Interestingness measures for association rule mining (ARM)**: Over the past two decades, ARM has received significant attention from both the data mining and machine learning communities. While data mining researchers focus on designing efficient algorithms to mine rules from large datasets, the learning community has explored applications of rule mining to classification. A major problem with rule mining algorithms is the explosion of rules even for moderate sized datasets making it extremely difficult for end users to identify both statistically significant and potentially novel rules that could lead to interesting new insights and hypotheses. Researchers have proposed many domain independent interestingness measures using which, one can rank the rules and potentially glean useful rules from the top ranked ones. However, these measures have not been fully explored for rule mining in clinical datasets owing to the relatively large sizes of the datasets often encountered in healthcare. Additionally, limited access to domain experts creates another obstacle for the review/analysis.

  In the first part of this dissertation, using an EMR dataset of 3.25 million visits to UKHealthcare clinics, we studied the trade-off between rule novelty and statisti-

cal significance using dozens of interestingness measures proposed in the literature and also a few additional measures we devised for this study. The rules we studied are of the form $E \Rightarrow Y$ where E (antecedent) and Y (consequent) are item sets formed from unique clinical variables: diagnoses, medications, procedures, and labs. Typically (including in our analysis), Y is a singleton and in biomedicine it is set to a medical/mental condition of interest. Here, we limited our analysis to only medications and diagnoses in our initial work and our consequent of choice is *depressive disorders*. The assessment of novelty of rules is conducted by a practicing psychiatrist (Dr. Rayapati) of UKHealthcare. Our results not only surface new interesting associations for depressive disorders but also indicate classes of interestingness measures that weight rule novelty and statistical strength in contrasting ways, offering new insights for end users in identifying interesting rules. The details of the methods used and results of this work are published in ACM BCB 2016 (Abar et al., 2016) and are presented in Chapter 3.

- **Toward causal association rule (CAR) mining**: Although association rules of the form $E \Rightarrow Y$ are interesting for further exploration, they may only indicate correlations that may be spurious. In fields such as biomedicine, it is more meaningful to identify causal association rules, where the antecedent E can be thought of as an event (e.g., taking a medication) that maybe *causing* the consequent Y (e.g., a condition, potentially as a side effect). Causal inference in biomedicine is a highly nuanced subject but most definitions of causality have a set of shared guidelines formulated by English epidemiologist Austin Bradford Hill in 1965. An obvious guideline is temporal precedence of E with respect of Y. So mining at patient-level LEMRs (the chronological EMR history) is a prerequisite. Additionally, the association must hold after accounting for potential confounding attributes, variables that may influence E and Y leading to a spurious correlation between the pair. Other guidelines involve establishing/identifying the existence of an actual plausible biomedical mechanism through which the causal association manifests and evidence through experiments (e.g., randomized controlled trials). Given our sandbox is the retrospective observational data in EMRs, we cannot account for all requirements needed for causality. But, we can account for confounders and ensure temporal precedence. With circumspection, we hence qualify our effort as moving "toward" CAR mining. Intuitively, however, CARs form a smaller and manageable high confidence hypothesis space compared with the full set of associations curated from EMRs.

A major challenge for CAR mining even in the retrospective setup is that all confounders are not known ahead of time besides the common ones such as gender, age, and race. Furthermore, approaches that rely on learning Bayesian networks (BNs) automatically from data do not scale to datasets with large variable spaces such as those encountered here. CAR mining is a more practical alternative for observational studies but has not been fully explored in current literature. In this dissertation, using an LEMR dataset of over 900,000 patient records with more than four million patient visits to the UKY medical center and affiliated clinics, we studied the causal effects involving diagnoses and medications, and patient demographics on two mental conditions: depressive disorders and anxiety disorders. The process of generating CARs starts with computational confounder detection. Next, using these confounders, we generate the so called "fair" datasets (FDs) that consist of matched LEMR pairs to calculate the strength of possible CAs. Then, rules with 95% CI-based odds-ratio lower bounds (ORLBs) greater than one, as calculated from the FDs, are selected as CARs. Finally, we compare these ORLB based causality rankings against expert judgments which are obtained from two practicing psychiatrists to assess the utility of our method. We identify interesting CARs and also find that the causality ratings produced by our method align with those assigned by the domain experts. Full details of this effort and corresponding findings are presented in Chapter 4.

- **Predictive modeling with contrast patterns and neural networks:** One in four Americans and three in four Americans aged 65 and older suffer from multiple chronic conditions leading to 71% of total healthcare spending in the U.S. Both in terms of sheer suffering and high out-of-pocket expenses, patients with multiple chronic conditions form a subgroup that are deservedly getting attention from the research and health services communities. Once diagnosed with such conditions, patients are usually on corresponding medications for the rest of their lives. Also, multiple chronic comorbidities greatly compromise immunity and can make people highly susceptible to acute infections that can rapidly lead to multiple organ failure. Thus, being able to predict such conditions well before they manifest fully can be of immense preventative and interventional value, which is clearly something all stakeholders — patients, doctors, healthcare facilities, insurance providers, and health policymakers — can get behind.

  Because LEMRs represent patient trajectories through the healthcare system, we can use supervised machine learning methods that take as input a prefix of a pa-

tient's LEMR and predict their future conditions. This can be further simplified, if we formulate it as estimating the probability of first diagnosis of a particular chronic condition in the future, predicted at different time horizons. In this dissertation, this is the line of work pursued for depressive disorders as the target condition group using different neural-network based representations of LEMRs. Unlike prior efforts that only employ recurrent neural network (RNN) variants, we also employ sequential *contrast patterns* (SCPs) derived from LEMRs and use RNN based sequence compositions on top of SCPs. By carefully varying washout periods, future time horizons, maximum inter-visit gaps, and minimum numbers of visits, we first examine the performance of existing LEMR based predictive modeling methods for different variants of LEMR input encoding. Next, we choose a particular cohort of patients (diagnosed with depressive disorders) with a washout period of six months, a one year maximum inter-visit gap, and with at least 30 visits made to UKHealthCare facilities. Using this data, we predict the first diagnosis of depressive disorders with a one year time horizon. We apply our novel modeling approach that hierarchically composes SCPs and SCP sequences using RNNs. Our results show that a hybrid model that combines a conventional RNN with our SCP-based RNN produces the best predictive performance. These models, their variants, and results are presented in Chapter 5.

Chapter 2 Related Work and Background

In this chapter, we will describe background concepts that are essential to the rest of the dissertation.

2.1 Healthcare Cost and Utilization Project

The HCUP (Healthcare Cost and Utilization Project, n.d.) aims to generate healthcare related datasets and software tools to build national data resources. Diagnosis and procedure classification via HCUP is carried out through its clinical classification software (CCS), which is used to group various diagnosis codes in this dissertation. HCUP has been created by the Agency for Healthcare Research and Quality (AHRQ), a federal organization that oversees and guides health services and care delivery aspects at the national level. The purpose of developing CSS is to generate clinically

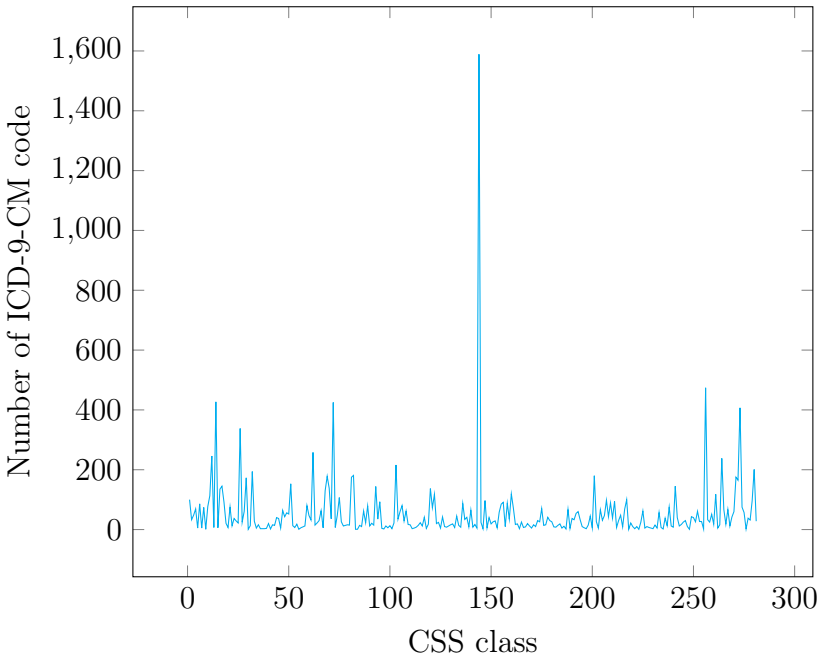


Figure 2.1: Number of ICD-9-CM codes for each CSS classes

meaningful classification of ICD and CPT codes. Instead of using a single code, using a group of codes related to same diagnosis or procedure is more beneficial in statistical studies to obtain high level estimates as opposed to estimates for all suble variants created originally for billing purposes. For example, ICD-9-CM consists of more than

14,000 diagnosis codes and these codes can be classified into 282 different classes using CSS' classification system. Figure 2.1 shows the number of ICD-9-CM codes in each CSS classe. While some classes are small, some have a large number of ICD-9-CM codes. For instance, class 144 (multi-level CSS category 18) "Residual codes; unclassified; all E codes" has 1589 different ICD-9-CM codes which is the class with highest number of codes while class 55 (multi-level CSS category 5.3.2) "Oppositional defiant disorder" has exactly one code. Tables 2.1 and 2.2 presents ICD-9-CM codes and descriptions related to depressive disorders and anxiety disorders respectively.

Table 2.1: List of ICD-9-CM codes related to depressive disorders

ICD-9-CM code	Description
293.83	Mood disorder in conditions classified elsewhere
296.20	Major depressive affective disorder, single episode, unspecified
296.21	Major depressive affective disorder, single episode, mild
296.22	Major depressive affective disorder, single episode, moderate
296.23	Major depressive affective disorder, single episode, severe, without mention of psychotic behavior
296.24	Major depressive affective disorder, single episode, severe, specified as with psychotic behavior
296.25	Major depressive affective disorder, single episode, in partial or unspecified remission
296.26	Major depressive affective disorder, single episode, in full remission
296.30	Major depressive affective disorder, recurrent episode, unspecified
296.31	Major depressive affective disorder, recurrent episode, mild
296.32	Major depressive affective disorder, recurrent episode, moderate
296.33	Major depressive affective disorder, recurrent episode, severe, without mention of psychotic behavior
296.34	Major depressive affective disorder, recurrent episode, severe, specified as with psychotic behavior
296.35	Major depressive affective disorder, recurrent episode, in partial or unspecified remission
296.36	Major depressive affective disorder, recurrent episode, in full remission
300.4	Dysthymic disorder
311	Depressive disorder, not elsewhere classified

2.2 Association Rule Mining

Let $\mathcal{I}$ be the union of all medications and diagnoses and any other biomedical variables that are of interest for each patient. For our purposes, a set $E = \{i_1, \ldots, i_k\} \subseteq \mathcal{I}$ is called a clinical *item set* with k items and a patient *visit transaction* $T = (pid, vid, I)$ is defined over $\mathcal{I}$ where *vid* is the patient visit ID, *pid* is the patient ID, and $I \subseteq \mathcal{I}$ is the item set corresponding to the current visit *vid*. The set of all visit transactions in

Table 2.2: List of ICD-9-CM codes related to anxiety disorders

ICD-9-CM code	Description
293.83	Mood disorder in conditions classified elsewhere
293.84	Anxiety disorder in conditions classified elsewhere
300.00	Anxiety state, unspecified
300.01	Panic disorder without agoraphobia
300.02	Generalized anxiety disorder
300.09	Other anxiety states
300.10	Hysteria, unspecified
300.20	Phobia, unspecified
300.21	Agoraphobia with panic disorder
300.22	Agoraphobia without mention of panic attacks
300.23	Social phobia
300.29	Other isolated or specific phobias
300.3	Obsessive-compulsive disorders
300.5	Neurasthenia
300.89	Other somatoform disorders
300.9	Unspecified nonpsychotic mental disorder
308.0	Predominant disturbance of emotions
308.1	Predominant disturbance of consciousness
308.2	Predominant psychomotor disturbance
308.3	Other acute reactions to stress
308.4	Mixed disorders as reaction to stress
308.9	Unspecified acute reaction to stress
309.81	Posttraumatic stress disorder
313.0	Overanxious disorder specific to childhood and adolescence
313.1	Misery and unhappiness disorder specific to childhood and adolescence
313.21	Shyness disorder of childhood
313.22	Introverted disorder of childhood
313.3	Relationship problems specific to childhood and adolescence
313.82	Identity disorder of childhood or adolescence
313.83	Academic underachievement disorder of childhood or adolescence

a given database is denoted as the visit database $\mathcal{V}$. A visit transaction (pid, vid, I) is said to *support* an item set E if $E \subseteq I$ and the *support* of E in the database $\mathcal{V}$ is defined as:

$$support(E, \mathcal{V}) = |\{vid : (pid, vid, I) \in \mathcal{V}, E \subseteq I\}|. \tag{2.1}$$

An item set is deemed *frequent* if its support is greater than a given minimum support σ. Thus, the set of frequent item sets with respect to σ is defined as:

$$\mathcal{F}(\mathcal{V}, \sigma) = \{E : support(E, \mathcal{V}) \geq \sigma\}. \tag{2.2}$$

Next, an Association Rule (AR) is a rule of the form $E \Rightarrow Y$ where E and Y are item sets and $E \cap Y = \emptyset$. The *confidence* of an association rule $E \Rightarrow Y$ denoted by

$$conf(E \Rightarrow Y, \mathcal{V}) = \frac{support(E \cup Y)}{support(E)}, \tag{2.3}$$

models the probability $P(Y|E)$ and establishes the association of the consequent item set Y with the antecedent item set E. Beside minimum support for item sets, we can establish a minimum confidence γ for ARs and define a stronger notion of frequent and confident ARs over a visit database $\mathcal{V}$ as the set

$$\mathcal{R}(\mathcal{V}, \sigma, \gamma) = \{E \Rightarrow Y : E \cup Y \in \mathcal{F}(\mathcal{V}, \sigma), conf(E \Rightarrow Y) \geq \gamma\}, \tag{2.4}$$

which consists of confidence thresholded ARs obtained from frequent item sets.

2.3 Term Frequency-Inverse Document Frequency

Term Frequency - Inverse Document Frequency (TF-IDF) (Robertson, 2004) is a commonly used statistical method in information retrieval. This is simple yet powerful formula measures the relationship of the term with a given document in corpus. Higher TF-IDF value indicates that the term have strong relationship with a given document. TF-IDF consist of two numerical values; Term Frequency (TF) and Inverse document frequency (IDF). TF-IDF value of a term for a given document in a corpus is defined as:

$$TF - IDF(t_i, d, \mathcal{C}) = TF(t_i, d) \times IDF(t_i, \mathcal{C}) \tag{2.5}$$

First value, TF, calculated as the number of occurrence of a term in a specific document in other word frequency of the term. TF value is defined as:

$$TF(t_i, d) = frequency(t_i, d) \tag{2.6}$$

Second value, IDF, is used to distinguish more meaningful words in a document. For instance, articles "the", "a", and "an" tend to occur in every text document in the corpus even if it has no meaning by itself. The idea is to give higher values to the word occurs in less documents to increase the importance of the document specific words. The basic and most commonly used formula to calculate IDF for a document collection and a term, is defined as

$$IDF(t_i, \mathcal{C}) = \log\left(\frac{\mathcal{N}}{n_i}\right) \tag{2.7}$$

In order to calculate IDF of the term t_i in a corpus $\mathcal{C}$, in equation (2.7), n_i is the number of documents contains the term, t_i, where $\{t_i : d \in \mathcal{C}, t_i \in d\}$ and d corresponds to a document in $\mathcal{C}$ as well as $\mathcal{N} = |\mathcal{C}|$ is the total number of documents in the corpus.

2.4 Odds Ratio

Odds Ratio (OR) (Morris and Gardner, 1988) is a commonly used statistical measurement to calculate the association of the exposure, a disease or medication, and the outcome, a medical condition. Table 2.3 shows a 2×2 contingency table for an AR. In this table, calculating OR gives the association between antecedent and consequent where antecedent means a exposed disease or medication as well as consequent corresponds to an another medical condition to relate. OR of an AR is calculated as

$$OR(E \Rightarrow Y) = \frac{a \times d}{b \times c}, \tag{2.8}$$

where $E \Rightarrow Y$ is an AR and a, b, c, and d are the values appear in the Table 2.3. An association between the exposure and the outcome is decided according to the OR value calculated. There are three different possible values of the calculation. If $OR = 1$, there is no association between the exposure and the outcome because odds of the outcome for the exposure is equal to odds of the outcome for not exposure. If $OR > 1$, there is a positive association between the exposure and the outcome as well as the odds of the outcome significantly increases when it is exposed to the antecedent. If $OR < 1$, the odds of the outcome is lower for the exposure.

	Consequent		
Antecedent	Y	$\neg Y$	Total
E	a	b	a+b
$\neg E$	c	d	c+d
Total	a+c	b+d	n

Table 2.3: 2×2 Contingency Table for Rule $E \Rightarrow Y$

Calculating the OR is not always enough to ascertain the association between the antecedent and the consequent. Therefore, we will also calculate the Confidence Interval (CI) to estimate the precision of OR. In order to calculate CI, we will first calculate Standard Error (SE). The formula to calculate SE for $\ln(OR)$ is defined as:

$$SE(\ln(OR)) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}. \tag{2.9}$$

With this the OR lower bound (ORLB) and the OR upper bound (ORUB) are given

as:

$$ORLB(E \Rightarrow Y) = e^{\left(\ln(OR) - 1.96 \times SE(\ln(OR))\right)} \tag{2.10}$$

and

$$ORUB(E \Rightarrow Y) = e^{\left(\ln(OR) + 1.96 \times SE(\ln(OR))\right)} \tag{2.11}$$

Equation 2.10 and Equation 2.11 show how to apply SE value to OR after calculating SE shown in Equation 2.9. These formulas, ORLB and ORUB, are calculated for the 95% confident interval. One of the advantages of using SE other than strengthening association is to avoid misjudgment caused by very small a, b, c, or d values.

2.5 Inter-Rater Reliability Scores

Many studies in biomedical domain involve observational rating scores from multiple people to demonstrate robustness of the annotation and the corresponding computational methodology. When there is more than one rater, the assessment of inter-rater reliability (IRR) is required to measure the degree of consistency among raters (Hallgren, 2012). IRR score shows the degree of homogeneity or consensus between the ratings given by the raters. There are a number of statistical measures to assess IRR in the medical domain including Cohen's Kappa (McHugh, 2012; Cohen, 1960; Cohen, 1968), Spearman's Rho (Mukaka, 2012), and Gwet's AC1Wongpakaran et al., 2013; Gwet, 2014. We will start by creating a 2×2 contingency table as shown in Table 2.4 to explain how each measure is calculated.

Table 2.4: 2×2 Contingency table for the assessment of IRR

		Rater 1		
		+	−	Total
Rater 2	+	a	b	a+b
	−	c	d	c+d
	Total	a+c	b+d	n

One of the most frequently used statistics to assess IRR is Cohen's Kappa also called kappa statistic first introduced in (Cohen, 1960). Cohen's Kappa symbolized

by κ and ranges between -1 and +1. κ value is calculated as:

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)} \tag{2.12}$$

where

$$P(a) = \frac{a + d}{n}, \tag{2.13}$$

and

$$P(e) = \frac{\frac{(a+c)x(a+b)}{n} + \frac{(b+d)x(c+d)}{n}}{n} \tag{2.14}$$

Another measure to assess IRR, which is frequently used in biomedical domain is the Spearman's Rho (Spearman, 1904) symbolized by ρ or r_s. The value of ρ ranges from -1 to +1. In order to calculate ρ, first we rank each ratings separately from lowest to highest. Then, for each data pair rank differences are calculated as d_i. Finally, ρ is calculated as:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n^3 - n} \tag{2.15}$$

Gwet's AC1 is also used for the assessment of IRR in biomedical data and was first introduced in (Gwet et al., 2002). As with Cohen's kappa, Gwet's AC1 uses probabilistic measurements to calculate IRR score which ranges between -1 and +1. This score is introduced to overcome some issues with κ which can produce different IRR values for same percent of agreement level. Additionally, κ is hard to interpret and unstable. Gwet's AC1 score is calculated as:

$$Gwet's\ AC1 = \frac{P(a) - P(\gamma)}{1 - P(\gamma)} \tag{2.16}$$

where P(a) is calculated as in Equation 2.13 and $P(\gamma)$ is calculated as:

$$P(\gamma) = 2q(1 - q) \tag{2.17}$$

with

$$q = \frac{(a + c) + (a + b)}{2n}. \tag{2.18}$$

According to the value calculated by an IRR measurement, 1 means that the degree of consistency among the raters is perfect while below and equal to 0 mean that there is no agreement. Level of agreement can be classified in six groups which are shown in Table 2.5.

Table 2.5: The agreement level of IRR measures

Value	Agreement Level
≤ 0	No Agreement
0.01-0.20	None to Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost Perfect

2.6 Sequential Contrast Pattern Mining

Contrast pattern mining (CPM) was first introduced by Dong and Li (1999) to find patterns contrasting multiple datasets or classes within a dataset. The idea is to find patterns that feature more prominently in one class when compared with other classes. For classification, we choose to compare two classes in an EMR dataset where positive class includes patients with user defined target condition and negatives class includes the rest of the patients. To apply contrast pattern mining approaches, let $\mathcal{I}$ be the union of all medical codes that can be used for patients. For our purposes, a pair w_j consists of a set of items and visit order index $w_j = \{(i_1, i_2, ..., i_l), t_j\}$ where $\{i_1, i_2, ..., i_l\} \subseteq \mathcal{I}$ for $l > 0$ and a set $W = \{w_1, w_2, ...w_j, ..., w_{k-1}, w_k\}$ is an ordered list item sets where $t_{k-1} < t_k$, which includes k pairs from k different patient visits. The set of all medical transactions in a given database is denoted as a sequential EMR database. The positive sequential database, EMR_p, is that includes all transactions that contain z and the negative sequential database ,EMR_n, is that includes all transactions missing z where z is the desired medical code that we are using as output. In this study, we only consider medical items that occur before z, and if there are multiple occurrences of z with m days difference, we will use the first occurrence of that output code. If there are multiple occurrences of z and there are more than m days between two occurrences, we create a new transaction using the medical items occurring with z in the same visit and after as a new transaction.

After creating two databases, we start mining all SCPs. Here, we consider 2 conditions: frequency and relative risk. A set $s = \{s_1, s_2, ..., s_m\}$ is m SP for $m > 0$ includes m items each $s_m \in \mathcal{I}$ where $s_{m-1} \in w_i$, $s_m \in w_j$, $i < j$, and $j - 1 > \alpha$, . Support of s defined as:

$$support(s, EMR_p) = |\{pid : (pid, W) \in EMR_p, s \in \mathcal{I}\}| \tag{2.19}$$

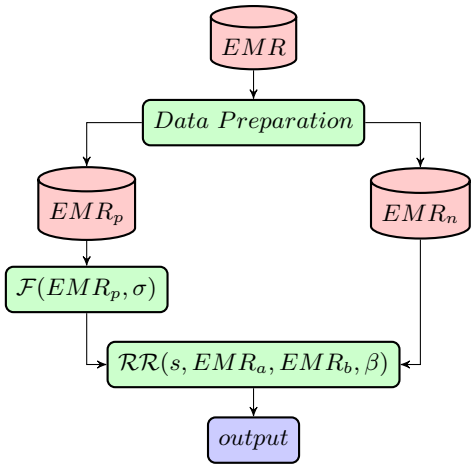


Figure 2.2: Sequential Contrast Pattern Mining Scheme

A pattern is frequent if its support is greater than a given support threshold σ for EMR_b or EMR_a. Hence, the frequency condition for a sequence s with respect to σ is defined as:

$$\mathcal{F}(EMR_p, \sigma) = \{s : support(s, EMR_p) \geq \sigma\} \tag{2.20}$$

To find all contrast patterns we first calculate relative risk as:

$$\mathcal{RR}(s, EMR_p, EMR_b) = \frac{support(s, EMR_p)/|EMR_p|}{support(s, EMR_n)/|EMR_n|}. \tag{2.21}$$

To find all SCPs, we will extract patterns which satisfy frequency condition for the positive dataset and then a separate relative risk condition: we consider SPs which are β times more likely to feature in the positive class. Therefore, all SCPs with respect to the relative risk condition are defined as:

$$\begin{aligned} SCP(\mathcal{F}(EMR_p, \sigma), EMR_b, \beta) = \\ \{s : s \in \mathcal{F}(EMR_p, \sigma) \ \& \ \mathcal{RR}(s, EMR_p, EMR_n) \geq \beta\} \end{aligned} \tag{2.22}$$

where β is the relative risk threshold for a sequence.

2.7 Recurrent Neural Network (RNN)

When sequential data is used, the order of each item in the sequence is important. For instance, we read sentences word by word in order to make sense of it and if we change the order, the sentence might become unintelligible or actually mean new things that the original sentence does not communicate. An RNN is a neural network with cyclical connections that naturally composes sequential information. Here, hidden layer of the

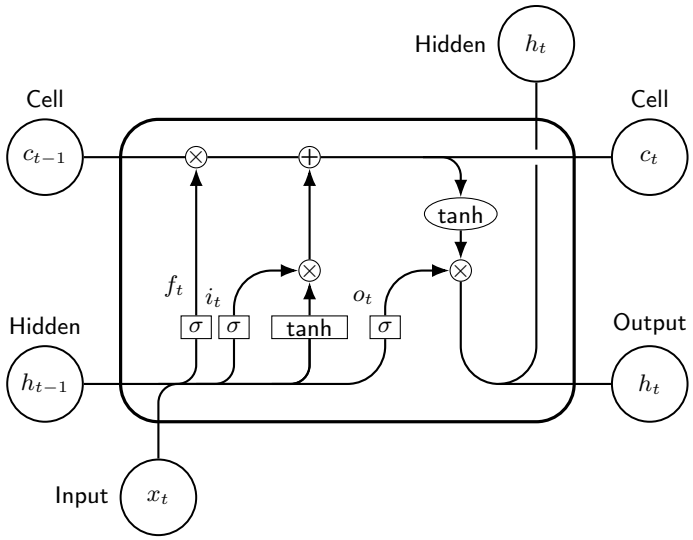


Figure 2.3: LSTM structure

current state is calculated by using the previous hidden layer and the current input and this value is the summary of the information given until that point. Formula to calculate the hidden layer of the current state is given as:

$$h_t = tanh(Wx_t + b_x + Uh_{t-1} + b_h) \tag{2.23}$$

where $U \in \mathbb{R}^{p \times p}$ and $W \in \mathbb{R}^{p \times m}$ are the parameters matrices, while $b_{\{x,h\}} \in \mathbb{R}^p$ are the related bias vectors with m being the embedding dimension of each word and p being hidden layer size. Here x_t is the input vector for current word (or any element in the sequence) and h_t is the hidden layer output of current state, and h_{t-1} is the hidden layer representation from the previous time step. There are two important variants of RNN used in the deep learning (DL) field.

2.7.1 Vanilla Long Short Term Memory (V-LSTM)

One well known RNN variant we employ in our studies is a standard LSTM, which we hencerforth term V-LSTM. The hidden unit of a V-LSTM model contains an input gate, output gate, and forget gate as shown in Figure 2.3. These gates have a value ranging from 0 and 1 and each gate has a specific role to improve the performance. Input gate controls how much of the new data will be used in current cell, forget gate decides what portion of the previous cell state is needed to be retained, and finally output gate decides how much information from current output will be sent to the next cell. More formally, a V-LSTM is specified as:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{2.24}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{2.25}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{2.26}$$

$$\tilde{c_t} = tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{2.27}$$

$$c_t = \sigma(f_t \odot c_{t-1} + i_t \odot \tilde{c_t}) \tag{2.28}$$

$$h_t = tanh(c_t) \odot o_t \tag{2.29}$$

Here, Equations 2.24, 2.25, and 2.26 show the input, forget, and output functions respectively. Equation 2.27 and 2.28 show memory cell equations using input and forget gate. Finally, equation 2.29 calculates the hidden layer of the current V-LSTM unit. i_t, f_t, and o_t are the input, forget, and output gates respectively. For embedding dimension of m, and hidden layer size p, $U_{\{i,f,o,c\}} \in \mathbb{R}^{p \times p}$ and $W_{\{i,f,o,c\}} \in \mathbb{R}^{p \times m}$ are the parameter matrices while $b_{\{i,f,o,c\}} \in \mathbb{R}^p$ are the related bias vectors. Also, $\sigma()$ is the sigmoid function and $tanh()$ is the hyperbolic tangent function. Here, x_t is the input vector, c_t is the memory cell, and h_t is the hidden layer of related LSTM unit. Our intuition in the context of LEMRs is to represented all structured codes just as one would embed words in a sentence. However, at each time step, there is an entire EMR instead of a single word. To handle this, an EMR can be represented by the simple average of embeddings of all constituent codes. Alternatively, different classes of codes (e.g., diagnoses, medications) can be averaged separately and then concatenated to come up with the fix dimensional embedding for an EMR. This latter approach leads to longer representations due to the separation of different types of codes.

2.7.2 Gated Recurrent Unit (GRU)

Another frequently used variation of the RNN model we employ is the GRU, which contains 2 gates: reset gate and update gate. This model provides comparable prediction power while the structure is simpler than the LSTM model. Figure 2.4 shows

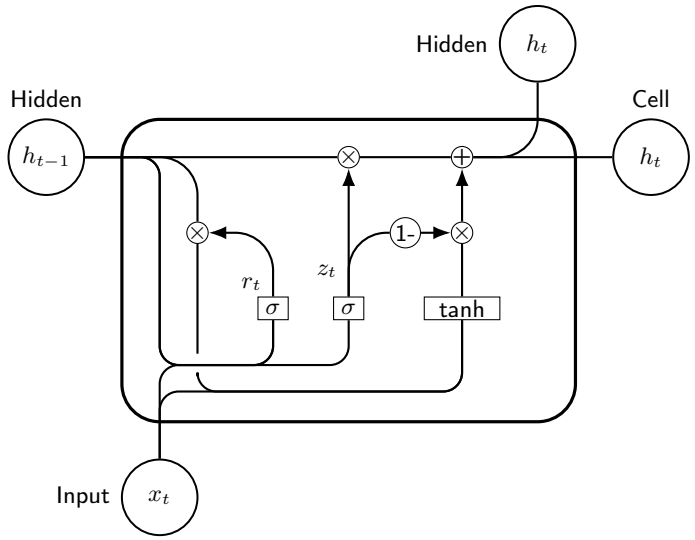


Figure 2.4: GRU structure

the structure of GRU and the formal description is as follows:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{2.30}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{2.31}$$

$$\tilde{h_t} = tanh(W_h x_t + r_t \odot U_h h_{t-1} + b_h) \tag{2.32}$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h_t} \tag{2.33}$$

Equations 2.30, 2.31, 2.32, and 2.33 show how to calculate the reset gate(r), the update gate (z), the intermediate memory unit ($\tilde{h}$), and the hidden layer output (h) respectively, where $U_{\{i,f,o,c\}}$ and $W_{\{i,f,o,c\}}$ are the parameter matrices while $b_{\{i,f,o,c\}}$ are the related bias vectors. The purpose of training is to learn these matrices and bias vectors. Because of the simpler architecture, GRUs tend to be faster. The representation of EMRs to be processed by GRUs is the same as explained for LSTMs in the previous section.

Chapter 3 On Interestingness Measures for Mining Statistically Significant and Novel Clinical Associations from EMRs

Association rule mining has received significant attention from both the data mining and machine learning communities. While data mining researchers focus more on designing efficient algorithms to mine rules from large datasets, the learning community has explored applications of rule mining to classification. A major problem with rule mining algorithms is the explosion of rules even for moderate sized datasets making it very difficult for end users to identify both statistically significant and potentially novel rules that could lead to interesting new insights and hypotheses. Researchers have proposed many domain independent interestingness measures using which, one can rank the rules and potentially glean useful rules from the top ranked ones. However, these measures have not been fully explored for rule mining in clinical datasets owing to the relatively large sizes of the datasets often encountered in healthcare and also due to limited access to domain experts for review/analysis. For this chapter, using an EMR dataset of diagnoses and medications from over three million patient visits to the University of Kentucky medical center and affiliated clinics, we conduct a thorough evaluation of dozens of interestingness measures proposed in data mining literature, including some new composite measures. Using cumulative relevance metrics from information retrieval, we compare these interestingness measures against human judgments obtained from a practicing psychiatrist for association rules involving the *depressive disorders* class as the consequent.

3.1 Introduction

Association rule mining (ARM (Agrawal and Srikant, 1994)) has emerged as an important methodology to gain insights into large databases of transactions each of which contains a set of items. ARM first gained popularity for market-basket analysis where each transaction consists of a set of products purchased by a customer. Using ARM, rules of the form $E \Rightarrow Y$ are extracted which indicate that a customer that buys a set of items E "tends" to buy items in Y in the same visit. ARs obtained for this domain have been used to better design product placement layouts in stores that encourage so called cross-selling among customers. Similar strategies are also being employed by online stores to dynamically generate product recommendations based on prior browsing/purchasing history. In the context of biomedicine

and healthcare, ARM has also been applied to EMR data for association analysis among biomedical and clinical variables (Brossette et al., 1998; Ordonez et al., 2006; Wright et al., 2010; Wright et al., 2013). Before we proceed further, we establish some primitives for ARM starting with the notion of a clinical item set.

3.1.1 Notions of Statistical Strength, Novelty, & Interestingness

Statistical significance and novelty are two important and complementary notions that make a rule desirable for further examination. Generally speaking, an AR is deemed statistically significant if its manifestation is not due to random chance. Statistical strength is a measure-specific notion that attributes a gradation or degree to the significance of the rule. Thus, we would at least want a rule to be statistically significant and also prefer for it to have high statistical strength. However, statistically significant ARs may not be meaningful or clinically relevant; even in cases when they are meaningful, they might be too obvious. For example, in our experiments, the association of antidepressants with depressive disorders is statistically significant but is very obvious to most end users. For ARM, the notion of novelty indicates the level of unexpectedness, surprise, or peculiarity associated with a rule. For example, the association between antidepressants and depressive disorders is considered not novel. For our current effort, to keep the terminology simple, novelty implicitly also includes the notion of clinical relevance or plausibility. In data mining literature (Geng and Hamilton, 2006; Shaharanee et al., 2011; Tan et al., 2002; Webb and Vreeken, 2014), "interestingness" has been used as an umbrella term to describe a combination of desirable rule properties including statistical strength and novelty and we employ the same usage for the rest of our chapter. Although novelty is sometimes considered a subjective measure, in this chapter we assess how various interestingness measures model novelty. Next we outline our main contributions.

3.1.2 Our Contributions

Prior results on applying ARM to clinical datasets (Wright et al., 2010; Wright et al., 2013) offer important insights but are based on relatively smaller datasets with a focus on rediscovering known associations already recorded in external knowledge bases. Hence they do not directly assess the novelty of the associations found. Furthermore, their evaluations consider only few interestingness measures (up to five) in their experiments and also limit the antecedent of an association to be a singleton. In our current effort

1. We use a dataset of diagnoses and medications from over 3 million patient visits to the UKY medical center and its affiliated clinics to obtain all ARs with singleton consequents and having minimum support 100 and minimum confidence 10%. We do not limit the rule antecedents to be singletons; they can be combinations of both diagnoses and medications.

2. We rank the specific set of rules with *depressive disorders* as the consequent using over 40 different interestingness measures including most measures introduced in data mining literature (Geng and Hamilton, 2006) and a few new measures we introduce in this chapter.

3. We obtain manually assigned novelty scores (1 – 5) for the set of rules in the union of top 100 rules from rankings produced by all interesting measures using the help of a practicing psychiatrist (Dr. Rayapati, the domain expert of this effort). We *combine* these novelty scores and odds ratio lower bounds (from 95% confidence intervals) for these rules to compare against all interestingness measures and identify classes of measures that trade-off novelty and statistical strength in contrasting ways. We also discuss the clinical plausibility of several novel associations identified in our analysis.

The central premise for all our work is to pick specific diseases of interest as consequents and identify groups of medications and other conditions (as antecedents) that are associated with them. The associations may themselves manifest due to comorbidity situations (if antecedents are diseases). They can be indicative of treatment relations or side-effect/adverse-reaction scenarios (if the antecedents are medications). Combinations of medications and diseases as antecedents can represent more nuanced and specific scenarios with high statistical strength.

3.2 AR Mining from Visits Data

ARM has been explained in Section 2.2. From a biomedical perspective, we can filter ARs $\mathcal{R}(\mathcal{V}, \sigma, \gamma)$ choosing interesting and meaningful consequents Y. For example, we can set $Y = \{\text{NSCLC}\}$, that is, a consequent with just one item, NSCLC, which corresponds to patient visits that had a diagnosis code for NSCLC.

As their name indicates, ARs are essentially associations (or correlations) and do not indicate causality, although they have been known to manifest when there is a causal relationship. ARs are also used as starting points to arrive at potential causal relations (Hill, 1965) using additional retrospective analyses involving confounding

factors (not all of which maybe recorded in a clinical database) or additional prospective experiments such as randomized control trials (which may not be feasible in all situations) (Shadish et al., 2002). We emphasize that the scope of this chapter is assessing rule interestingness measures in the context of *ranking large AR sets to enable discovery of interesting associations* that can lead to novel hypotheses. Next we discuss the notions of statistical strength, novelty, and interestingness of rules generated by ARM. Here we primarily discuss the clinical dataset and methods used to extract ARs.

3.2.1 Clinical Dataset Used

Our dataset is extracted from all patient visits ($\approx$ 3.25 million) during the ten year period 2004-2013 to the UKY medical center and its affiliated clinics. Each visit transaction consists of medications and diagnoses recorded during a particular patient visit*. We also removed nearly 12,000 transactions that are very large (with 35 or more elements per visit). Although rare and in this case constituting only 0.3% of the full dataset, presence of such long transactions renders existing approaches to ARM impractical given they all rely on generating frequent item sets as an intermediate step. Thus we are still left with $\approx$ 3.25 million visits from around 572,000 unique patients. Thus, on average, each patient had about 5.66 visits during the decade. Given the ten year window of the study, we chose to treat different visits by the same patient as giving rise to different transactions. This way, the co-occurrences of medications and diagnoses are guaranteed to have the same time stamp in all our transactions.

The dataset has 11,877 unique ICD-9-CM codes and 1032 unique medication codes by *Cerner Multum*^{TM} *Lexicon Plus* codes which are also used by CDC for their medical care surveys. Current ARM approaches, even with the advent of "big data" approaches, do not scale well to thousands of unique items for patient visit databases with large transaction sizes especially if the minimum confidence and threshold are chosen to be small, which is critical to surface novel associations; high support and confidence rules may satisfy statistical strength requirements but tend to represent common knowledge for most end users. At lower thresholds, scalability issues mostly arise because of the combinatorial explosion of possible antecedent sets. Furthermore, considering all unique codes may not offer enough statistical strength (due to sparsity) or yield informative rules (for manual AR interpretation). For example,

*Although other variables such as procedures and labs are available, for computationally tractability we limited our current study to medications and diagnoses.

researchers might be more interested in knowing statistically significant and novel associations of penicillins with other conditions rather than be subjected to a deluge of weak associations involving specific penicillins such as Amoxicillin, Ampicillin, and Dicloxacillin. However, sparsity issues may be overcome by working with much larger datasets compared to the dataset used in our current effort.

Given above scenarios, we group diagnosis and medication codes using conventional approaches. For diagnoses, we use ICD-9 code classes (Healthcare Cost and Utilization Project, n.d.) developed by the HCUP, an affiliate of the AHRQ in the US Department of Health and Human Services. These classes group related codes resulting in 282 classes for the 11,877 codes in our dataset. For example, the HCUP class for *cancer of breast* groups 13 different ICD-9 codes covering all female breast cancer codes, male breast cancer codes, and a code for personal history of breast cancer. We rolled-up the Multum medication codes using their class hierarchy which resulted in 150 classes (e.g., Penicillins). In each transaction, we then replaced the codes with the corresponding HCUP and Multum classes resulting in a total of 432 unique items (HCUP and Multum classes) populating 3.25 million transactions.

3.2.2 Generating Association Rules

Although there are several efficient implementations that extract frequent item sets (Han et al., 2000; Zaki, 2000), including those that work on big datasets using MapReduce (Moens et al., 2013), for our purposes the LCM Ver. 3 by Uno et al. (2005) that exploits a clever combination of bitmaps, prefix trees, and array lists worked best. We used a minimum support $\sigma = 100$ and confidence $\gamma = 10\%$ for singleton consequent rule generation. That is, in each AR, we require that the antecedent items and consequent co-occur at least 100 times in over 3 million transactions and at least 10% of the transactions that contain the antecedent set also include the consequent. This is in line with other efforts (Wright et al., 2010; Wright et al., 2013) on applying ARM to clinical datasets. LCM generated nearly 22 million rules for our dataset.

At this point, to evaluate interestingness measures for both statistical strength and novelty, we needed to pick a narrow focus. According to the National Comorbidity Survey Replication (2001–2003), 68% of adults with mental disorders have medical conditions and 29% with medical conditions have mental disorders (Kessler et al., 2004). A February 2011 Robert Wood Johnson Foundation (RWJF) research synthesis report (Druss and Walker, n.d.) presents evidence that this subgroup of people with mental and medical disorder comorbidities are at significant risk for poor quality of care and high costs. Depressive disorders are one of the most common

mental disorders especially among adults and hence we picked the corresponding HCUP class for our focused study. The *depressive disorders* HCUP class has sixteen ICD-9 codes, which represent all variants of depression in . Our dataset has 54,923 transactions with a depressive disorder code. Post filtering all rules with depressive disorders as the consequent, we obtained 126,540 rules. Upon on manual observation, many of these rules had *antidepressants* as an element of the antecedent. Since the presence of this well known drug class that treats depression leads to uninteresting associations, we removed those rules with antidepressants as part of the antecedent, which resulted in 75,465 rules. These are the rules we ranked based on different interestingness measures.

3.3 Assessing Interestingness Measures for Association Rule (AR) Ranking

We ranked all the 75,465 rules with depressive disorders as the consequent class using nearly three dozen probability based objective interestingness measures from a recent survey by Geng and Hamilton (Geng and Hamilton, 2006, Table IV). This list includes popular measures such as confidence, lift, conviction, odds ratio, and information gain. Additionally, we added the χ^2-measure as it is well known for studying statistically significant associations (Hämäläinen, 2011; Wright et al., 2010). We also introduced some new measures which we describe here.

3.3.1 Additional Interestingness Measures

To model novelty, we introduce the notion of AIRF for a given AR $E \Rightarrow Y$. Recall from Chapter 2, $\mathcal{R}(\mathcal{V}, \sigma, \gamma)$ represents the set of ARs for the visit databases $\mathcal{V}$ satisfying minimum support σ and confidence γ. Let $\mathcal{R}^Y \subseteq \mathcal{R}(\mathcal{V}, \sigma, \gamma)$ be the set of rules with Y as the consequent from the full set of rules, assuming the database $\mathcal{V}$, σ, and γ are fixed. We define

$$AIRF(E \Rightarrow Y) = \frac{\sum_{x \in E} \frac{|\mathcal{R}^Y|}{|\{R:\ R \in \mathcal{R}^Y \ \wedge \ x \, is \, in \, antecedent \, of \, R\}|}}{|E|}.$$

Inverse rule frequency is analogous to inverse document frequency (IDF) in the TF-IDF term weighting scheme popular in information retrieval. The higher the AIRF of a rule $E \Rightarrow Y$, the fewer are the rules that contain elements of E as part of their antecedents – in this sense, rules with higher AIRF are expected to be

novel/peculiar. The rationale for AIRF follows from the justification for IDF (Robertson, 2004).

Odds ratio (OR) is a well known measure for studying associations in epidemiology[†] and more specifically, the ORLB (Morris and Gardner, 1988) of the 95% confidence interval around sample OR is used as an important measure for assessing statistical significance or lack thereof. $ORLB > 1$ indicates a statistically significant association with higher values indicating stronger associations. Our new measures of interestingness for a rule $E \Rightarrow Y$ include its $AIRF$, $ORLB$,

$$\frac{ORLB}{\log_2(|E| + |Y|)}, \ \textit{and} \ \frac{AIRF \cdot ORLB}{\log_2(|E| + |Y|)}, \tag{3.1}$$

where $\log_2(|E| + |Y|)$ indicates the length of the rule. (Note $|Y| = 1$ for our purposes and the expression equals 1 for singleton associations where additionally $|E| = 1$). Given ORLB indicates statistical strength and AIRF models novelty, we combined both in the product measure. Although we support longer rules with $|E| > 1$, very long rules are not interesting as they capture highly specific scenarios that are not amenable to reasonable interpretation and typically have low support as noted in prior efforts (Hämäläinen, 2011). At the same time we do not want to severely discount long rules. So to prefer smaller rules and dampen the effect of the length on overall interestingness score, we use $\log_2(|E| + |Y|)$ in the denominator of the two measures in equation 3.1.

3.3.2 Domain Expert Novelty Assessments

We used ORLB introduced in Section 3.3.1 as a proxy for statistical strength in our final assessment of all interestingness measures given it is routinely considered in biostatistics. The rationale for using ORLB over OR is that ORLB balances assurance that the result is not due to chance with the strength of the estimated effect, considering the variance of the estimator. However, we do not have a similar measure for novelty. Given it is unrealistic to have domain expert assessments on 75,000 rules we combined the top 100 rules from each of the rankings produced by all interesting measures discussed in this section. That is, given $\mathcal{M}$ is the set of all

[†]For prospective studies, relative risk (RR) is a more intuitive measure of association strength, but OR is a symmetric measure that is typically used for retrospective studies and approximates RR for rare outcomes (Rosner, 2015, Chapter 13.3). The advantage of OR over RR is that OR can be validly estimated whether random samples are drawn from the population as a whole, from exposure/risk factor strata, or from outcome strata.

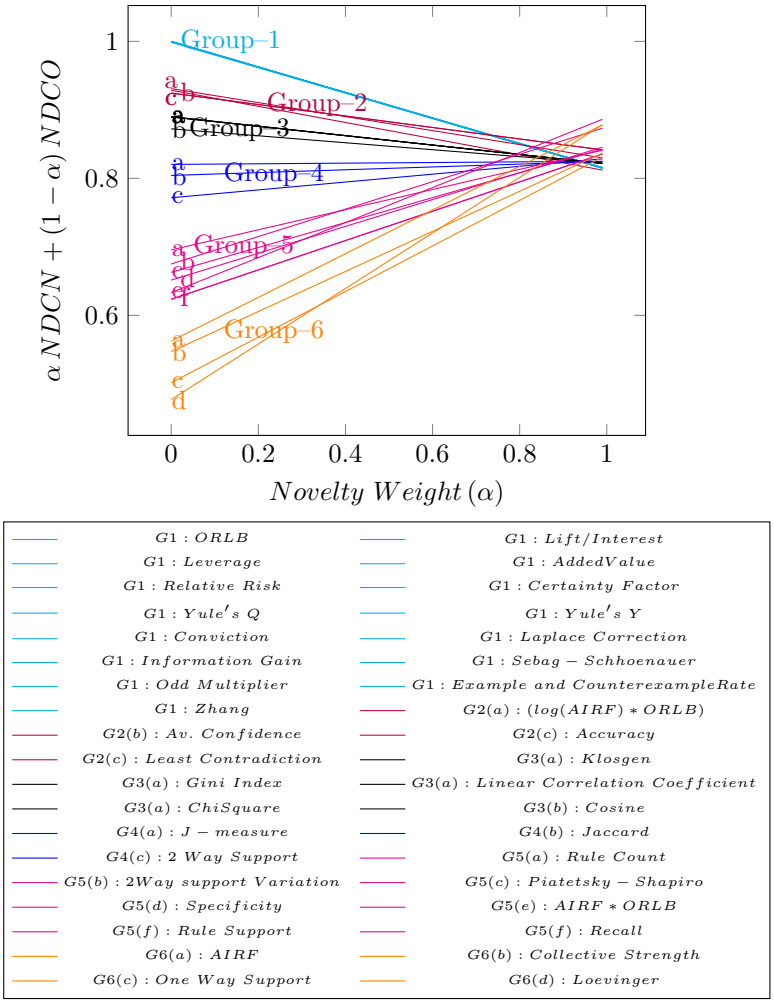


Figure 3.1: Interestingness measure profiles with novelty-statistical strength trade-offs

interestingness measures, human annotations are assigned to the set of rules

$$\bigcup_{m \in \mathcal{M}} Rank_m^{100}(\mathcal{R}^Y), \tag{3.2}$$

where $Rank_m^k$ indicates a function that returns the top k rules (without any limitations on rule length) obtained by ranking using measure m. In addition to this, all singleton antecedents which had an $ORLB > 1$ were also presented to the domain expert. We did this because singleton associations ($|E| = 1$) are easier to interpret, relatively very few compared to longer rules, and $ORLB > 1$ already indicates statistically significant association.

Novelty ratings were assigned on a scale of 1 to 5 (with 5 indicating most novelty) by a practicing psychiatrist from the university's department of psychiatry. As we indicated earlier in Section 3.1.1, the notion of novelty (regardless of the degree) for

our purposes includes plausibility. So a rating of 1 for a rule indicates it is a well-known association whose underlying mechanism is also reasonably understood. On the other hand a rating of 5 means it is a highly novel rule that is also clinically plausible although the details of the mechanism may not be as clear as for a rule with rating 1. This is to be expected given high novelty usually also implies that pertinent broad knowledge is lacking (see Section 3.4 for literature search based evidence for this). The assessments are informed by the physician's general medical knowledge and experiences as a practicing psychiatrist. Besides the actual rules, no other information was provided to the physician, who was requested to provide additional qualitative feedback on associations that were deemed highly novel. We chose the top 100 rule set union from all measures in the interest of domain expert time needed for novelty assessment. This limit has resulted in over 550 rules and we believe choosing larger thresholds could help for future efforts.

3.3.3 Comparison of Interestingness Measures

Next we compare interestingness measures discussed in this section across two dimensions, statistical strength and novelty, using rule ORLBs and psychiatrist assigned novelty scores as corresponding proxies, respectively. Using each interestingness measure, we rank all rules in equation 3.2 and any other singletons with an $ORLB > 1$ for the depressive disorders consequent. For reviewing convenience for the domain expert and subsequent analysis, we split all these rules into singleton and non-singleton antecedent rules. We ended up with a total of 231 singleton rules and 334 non-singleton rules each of which was assigned a novelty score (1–5).

The NDCG (Järvelin and Kekäläinen, 2002, Sections 2.2–2.3) is a popular rank quality metric in information retrieval (IR). It is typically used for search engines to measure the *gain* in terms of graded relevance of retrieved documents where relevant documents higher up in the ranking are given more weight compared with those that come later in the ranking. For interestingness measure comparison in our effort, we adapt NDCG to suit our purposes and compute normalized discounted cumulative novelty (NDCN) (from expert assigned scores) and normalized discounted cumulative ORLB (NDCO) based on the rule ranking produced according to each measure. Instead of the relevance judgment score of a retrieved document, we used a rule's novelty score (for NDCN) and ORLB (for NDCO). Besides this replacement of relevance scores with novelty and ORLB values, the exact expression used for NDCN and NDCO is identical to that of NDCG (Järvelin and Kekäläinen, 2002, Equation (2)). We then sorted all measures based on the corresponding NDCN and NDCO values to

identify best measures from the perspective of novelty and statistical strength. Like NDCG, the normalization aspect of our formulations implies both NDCN and NDCO take values in $[0, 1]$, where a value to closer to 1 indicates higher rank quality.

Instead of a single measure, we found classes of measures that scored similarly based on NDCN and NDCO values. Specifically, for singleton rules, $AIRF \cdot ORLB$ gave the highest NDCN value of 0.88. Measures such as $AIRF$, Loevinger, and 2-way support variation (Geng and Hamilton, 2006, Table IV) had NDCN values in $[0.87, 0.88]$. The lowest values for NDCN resulted from measures such as relative risk, Yule's Q, information gain, lift/interest, and conviction (Geng and Hamilton, 2006, Table IV) all of which had NDCN value around 0.81. On the other hand, for NDCO, these measures gave the maximum values of around 0.99. Similarly, Loevinger, which is among the top scorers for NDCN, generated the lowest NDCO score of 0.47. This demonstrates the clear trade-off between statistical strength and novelty in terms of what several interestingness measures are trying to capture.

To further compare the measures where different levels of importance are given to novelty (vs statistical strength), we plotted a combination metric

$$\alpha \cdot NDCN + (1 - \alpha) \cdot NDCO \quad \in [0, 1]$$

for all measures for $\alpha = 0, 0.01, 0.02, \ldots, 0.99, 1$. The results of this plot are shown in Figure 3.1. For convenience, we divided the measures into high level groups and appropriate subgroups with memberships as indicated in the legend of the figure. First we consider the six (Group1–6) different high level groups of measures with their corresponding performance profiles as α is varied. These groups were identified based on how they cluster together when statistical significance is solely considered (that is, when $\alpha = 0$). Group-1 has fifteen measures and is heavily biased toward maximizing statistical strength but also represents the top set of measures even when assigning equal importance to novelty and strength ($\alpha = 0.5$). Group-4's performance is relatively stable but does not generate superior overall performance. Group-5 archives novelty values that are higher than those of groups 1, 3, and 4. If we look at the measures from a novelty perspective, they break down into two distinct groups as can be observed when $\alpha = 1$ on the right most end of the plot in Figure 3.1. The first group achieves higher NDCN values and has four measures: $AIRF \cdot ORLB$ (G5(e)), $AIRF$ (G6(a)), Loevinger (G6(d)), and two way support variation (G5(b)). The rest of the measures can be clustered into the second group. Depending upon a particular user's specific preferences toward strength and novelty, he/she can choose

an appropriate measure based on variations noticed in the figure. When $\alpha = 0$, measures in Group-1 are recommended; but to maximize novelty ($\alpha = 1$), measure $AIRF \cdot ORLB$ appears superior.

For longer rules, the highest NDCN of 0.933 was achieved by $ORLB/\log_2(|E| + |Y|)$, where $|E|+|Y|$ represents the length of the rule. However, several other measures such as relative risk, Yule's Q, information gain, lift/interest, and conviction all had NDCN very close to 0.93. For NDCO, the highest value of 1 was achieved by Yule's Q and Yule's Y (besides ORLB). Other measures that scored well for NDCN also scored close to the maximum value for NDCO. Hence for longer rules, the trade-off effect that led to different groups of measures that lean toward either novelty or statistical strength does not seem to exist.

3.4 Quantitative & Qualitative Analysis of Novel Rules

We took two different approaches to analyze rules that were judged novel by the domain expert. We first manually mapped the medications and disease classes to MeSH terms, which are used to categorize biomedical articles by the US National Library of Medicine (NLM). Our visit item to MeSH mapping was done based on simple look-ups of the item names in the MeSH browser (`https://www.nlm.nih.gov/mesh/MBrowser.html`) and with the assistance of NLM's Unified Medical Language System (UMLS) to identify synonymous names. Since some HCUP and medication classes have multiple related items, some of them translated to multiple MeSH terms. MeSH terms are typically used to search biomedical articles using NLM's PubMed web application. For a given singleton rule $\{e\} \Rightarrow \{y\}$, we searched PubMed with the Boolean query

$$\left(\bigvee_{t1 \in MeSH(e)} t1 \right) \bigwedge \left(\bigvee_{t2 \in MeSH(y)} t2 \right)$$

for items e and y where $MeSH(x)$ denotes the MeSH term set for item x. For those singleton rules with expert assigned novelty scores ≤ 3 (total: 170), we retrieved an average of 1168 articles per rule, but the corresponding average over rules with novelty scores ≥ 4 (total: 61) is 264 and for those rules that have the top score five (total: 17), the average is 70 articles. This clearly shows that expert assigned scores seem to be aligned with what is reported in scientific literature based on co-occurrence analysis. For example, the drug class proton pump inhibitors (PPIs) has ORLB 9.98 and pulmonary heart disease has ORLB 3.07. Both were assigned a novelty score of 4 for their association with depression. For the corresponding conjunctive

queries with depression, one article was returned per query, but in both cases manual review of the articles revealed no explicit discussion of the associations. For PPIs, a similar association was found with myocardial infarction by Shah Shah et al. (2015) in a recent effort. Our findings regarding rheumatoid arthritis (ORLB: 2.37) and osteoarthritis (ORLB: 4.07) are also inline with a recent and thorough study (Ryu et al., 2016) that specifically looked into the impact of 24 chronic conditions on diagnosis of major depressive disorder, which differs in some aspects from the HCUP depressive disorders class used in our effort.

Table 3.1: Antecedents with novelty ≥ 4 and ORLB ≥ 5

Antecedent	Novelty	ORLB
CNS stimulants	5	7.65
Antianginal agents	5	7.21
Acute posthemorrhagic anemia	5	6.64
Endometriosis	5	6.46
Somatoform disorders	4	12.33
Antacids	4	9.82
ACE inhibitors	4	8.35
Anticoagulants	4	8.27
Hormonal antineoplastics	4	8.23
Esophageal disorders	4	8.02
Muscle relaxants	4	7.28
Antiplatelet Agents	4	6.77
Leukotriene modifiers	4	6.73
Immunostimulants	4	6.52
Quinolones	4	6.24

Next, based on direct inputs from the domain expert, we comment on the clinical plausibility of some of the high scoring (novelty score 4 or 5) associations for depressive disorders. Novel associations with depression are identified for conditions such as anemia (ORLB: 6.64), asthma (ORLB: 4.83), congestive heart failure (ORLB: 4.54), coronary atherosclerosis (ORLB: 3.75), and pulmonary heart disease. All these conditions can compromise oxygen flow to the brain and can contribute to microvascular

injury in white matter and contribute to atypical depression. Parkinson's disease (ORLB: 5.3) and migraine (ORLB: 3.48) affect the brain and their treatments will more than likely disrupt neurotransmitter systems implicated in depression. Behavioral disorders such as ADHD (ORLB: 8.1), oppositional defiant disorder (ORLB: 15.6), and conduct disorder (ORLB: 7.73) occur in the context of unclear biological vulnerability and psychological constructs of low self-esteem which tend to perpetuate social chaos similar to the individual's own developmental experience. Such social stress factors (poverty, unemployment, inconsistent employment, legal consequence, substance use, divorce, psychological trauma) have also been implicated in depressive disorders. So far in this section, we have looked at 13 singleton novel antecedents with some reflection on clinical relevance. In Table 3.1 we show the remaining novel (score ≥ 4) associations with ORLB ≥ 5.

There were a significant number of non-singleton associations with depression where the antecedent involves the suicide and intentional self-inflicted injury HCUP class along with other conditions and medications. For instance, the combination of the suicide HCUP code with osteoarthritis had ORLB over 150 but is peculiar and could be due to the observed but not thoroughly understood link between inflammation (conditions with the "itis" suffix) biomarkers and depression. Similarly, the association of suicide and alcohol related disorders with depression is well known but when epilepsy is added as a third condition to the antecedent, the association becomes statistically much stronger but also novel given seizures (from epilepsy) are considered therapeutic for mood disorders. Given seizures are also a complication in alcohol withdrawal, epilepsy might be indicating a more complex exacerbating alcohol related disorder.

3.5 Concluding Remarks

With innovations in computer science, informatics, and health information technology, EMR data from healthcare facilities and claims data from private and government sponsored insurance programs have become very rich sources for mining new insights for disease prevention and treatment. ARM has shown promise in other fields and is currently being actively explored for biomedicine to generate new hypotheses and also to build interpretable predictive models. An important concern in this era of big-data is dealing with vast number of rules output by ARM methods. In this chapter, we evaluate over 40 interestingness measures (including some new measures) for effective ranking of ARs across two desirable properties of statistical strength and novelty.

Using domain expert assigned novelty scores and ORLB for statistical strength, we adapted information retrieval metrics to assess various interestingness measures and identified classes of measures that seem to inherently weight novelty and statistical strength in contrasting ways. End users can utilize a particular class of measures depending on their goals that might influence their preferences for novelty and statistical strength. We conducted quantitive and qualitative analyses of some of the novel associations obtained as part of this effort. To our knowledge, this is the first effort to conduct a broad scoped comparative analysis of interestingness measures for clinical ARM involving subject matter expert driven novelty assessment.

Chapter 4 Toward Causal Association Rule Mining

Understanding the variation in risk profiles of conditions based on demographic attributes of a patient is a well-known activity to tailor treatments to subpopulations (McAlpine and Mechanic, 2000; Gove, 1984; Lasser et al., 2000). Researchers found evidence that a person's (ill)health condition is influenced by demographic information such as the age, gender, and race. Other life style related attributes such as BMI and smoking status are also known to affect health conditions. Demographic and life style variables may be mediators of a potentially spurious association between the antecedent and consequent of a conventional association rule (AR). As such, these variables are known to *confound* the real relationship between the entities in the AR. Thus it is critical to account for these confounder variables (simply called confounders) in assessing the strength of associations. The idea is to isolate a potential causal effect by appropriately considering other variables that may be exaggerating the relationship. This will help researchers identify new hypotheses to design interventions or recommend preventative measures. However, choosing demographics alone as confounding variables is not enough. There maybe be other intermediate conditions/medications that have a confounding effect on the participants in an AR. Hence it is also necessary to identify such variables and then account for them when computing the statistical strength of any AR to determine if it is causal or not. The ARs $E \Longrightarrow Y$ derived from LEMRs with the temporal precedence constraint that are deemed statistically significant after accounting for confounders are termed causal associations (CAs) in this chapter.

There have been different strategies to generate CAs using the medical history of patients. Randomized controlled trials (RCTs) are an experimental strategy of discovering CAs from randomly separated groups of cases and controls. The control group is given the standard treatment and the other group uses the new treatment whose efficacy is being evaluated. To find CAs, RCT method is an effective method among researchers who work in the medical domain. However, it may not be ethically viable to conduct an RCT for each and every possible association of interest. Also, not all associations are treatment related. Identifying causal side affects of long-term exposure to certain medications cannot be typically launched as a prospective RCT given the high risk of harm to the patients. Hence, researchers created alternative computational retrospective methods to obtain comparable results. One method is the causal Bayesian network (BN) model (Spirtes, 2010) which utilizes a graphical

model with nodes and edges where each edge shows a CA between connected two nodes, the variables. The major drawback of this method is that it is computationally expensive to generate the graph for datasets with a larger number of variables. For the itemset generated from our EMR database, creating such a large network is impractical. The need of creating an automatic method remains unfulfilled for larger datasets with BN. Recently, CAR mining method (Li et al., 2016) was developed to generate CAs using the advantages of association rule mining (ARM) approaches for the larger databases like the EMR database. In our work, we adapt this approach to the clinical setting to improve the domain expert based validity of generated CARs.

In Section 4.1, we explain essential related works and list our contributions. We present our EMR dataset and its details in Section 4.2. Additionally, in Section 4.3, we discuss our method to mine CARs from the EMR dataset. Then, the experimental configurations, results of our methods, causality ratings given by two medical experts, and the comparison of our method with the expert scores are explained in Section 4.4.

4.1 Related Works and Our Contributions

Detecting CAs in the health domain is a very important task and studied by different researchers (Moore et al., 2007; Abuse et al., 2006; Moore et al., 2007; Blakely et al., 2003). Many researchers identify a condition as well as possible causes of that particular condition and experiment to prove the association or the independence of variables. Taylor et al. (1999) conducted an epidemiological study to answer the question if autism is caused by a vaccine called MMR and showed that they are not causally associated. According to the causal study conducted by Gillison et al. (2000), human papillomaviruses and head & neck squamous cell carcinomas are causally associated.

In (Moore et al., 2007), researchers studied the causal effect of using cannabis, which is the most common illegal drug in the most of the countries (Abuse et al., 2006), on any psychotic outcome and found consistent results of increased risk. Besides, they mentioned that conducting RCTs for a study which includes the effect of an illegal drug is neither practical nor ethical. Smith and Ebrahim (2002) mentioned that employing an RCT study is hard to accomplish. Therefore, the successful attempts of observational studies are not always verified. According to (Wald et al., 2006), verifying the observational studies is not practical since a very large scale RCT is typically needed. Hence, researchers have been developing automatic CA extraction methods. BN (Heckerman et al., 2006; Spirtes, 2010) based approaches are provided

for automatic detection of CAs which are often computationally expensive. Hence, this method is not feasible for large numbers of variables. In order to overcome this problem and automatically extract CAs, a CAR mining approach is introduced by Li et al. (2016). This method is suitable for large datasets and is an efficient alternative of BNs for finding CAs.

In healthcare studies, researchers set out to identify the real causes of mental illnesses. These researchers found a particular condition or variable and checked whether it has a causal association with a specific mental illness. Moore et al. (2007) employed a longitudinal study to reveal the causal effect of cannabis usage to schizophrenia and depression. After reviewing patient data, they found sufficient evidence that using cannabis increases the risk of mental illnesses. In (Miech et al., 1999), the authors researched the CA between a demographic variable which is low socioeconomic status and four different mental disorders: anxiety disorders, depression disorders, antisocial disorders, and attention deficit disorders. Gariepy et al. (2010) conducted a literature review to clarify the association between obesity and anxiety disorders. They reviewed 16 epidemiological articles about obesity and anxiety disorders. Their result shows that there is a positive association between these two conditions.

Opstelten et al. (2006) conducted a study and used age as a confounding variable while assessing the relationship between the gender and herpes zoster. Consequently, they found the gender is an independent risk factor for herpes zoster in certain ages. Schneider et al. (2005) studied the effect of the gender and age groups on *"axis I disorders"* which includes some of the most common mental disorders such as anxiety disorders, eating disorders, and mood disorders. Some researchers used demographics as confounding variables while Blakely et al. (2003) used mental health conditions as confounding variables for investigating the effect of unemployment on suicide. Low et al. (2016) compared 18 different methods to generate high-dimensional confounders from the EMR database. After evaluating all the methods, they concluded lasso regressionTibshirani, 1996 as the best method. Therefore, we exploit lasso regression for our confounder list generation process.

Researchers in medical domain generally use their previous medical expertise to focus on a specific condition and a potential hypothesized causative agent of that specific condition. Then they conduct a study, which is an RCT in most cases, to verify the CA of the hypothesized variable. Hence, the approach used by researchers for a condition is not suitable for another condition. Moreover, the data they collect for a study is limited to itself. As far as we know, an automatic and robust approach, which is applied to a large dataset with expert verification is missing. In order to fill

this gap, we pursue a study with the following contributions.

- We use an EMR dataset of demographic attributes, diagnosis and medication codes from around 922,000 patients and 4.15 million patient visits to the University of Kentucky medical center and its affiliated clinics to obtain CARs for anxiety disorders and depressive disorders. We start with all statistically significant ARs having the minimum support of 100, the minimum confidence of 10% and odds ratio lower band (ORLB) (from 95% confidence interval) value of greater than one. We decided to employ singletons as exposure so that our medical experts can review all exposures manually in a timely manner.

- We impose temporal precedence since the cause of an event must precede the event which we term as *outcome* for this study. To approximate the identification of first diagnosis, we use a washout period of six months where we guarantee that each patient record we use has at least six months of known prior history in our dataset before a diagnosis of the outcome is made.

- We average ORLB values of 11 different criteria (confounder lists) where the confounders include demographics and those automatically generated by lasso logistic regression.

- We obtain manually assigned causality scores (1-5) for all ARs for depressive disorder and anxiety disorder outcomes using the help of two practicing psychiatrists (Dr. Rayapati and Dr. Zwiebel). Finally, we compare the results of our method with domain experts' results to assess the validity.

The main objective of this study is to apply CAR mining approach to our EMR database to automatically generate accurate and informative CARs for a chosen target condition. Finding such causes will help physicians to better understand the possible future diagnoses of a disease and to assess risk factors for it. Our approach will compare the outcome to all possible exposures. Therefore, it can surface a new relationship which is previously unknown by the physician.

4.2 Clinical Dataset Used

In this research, we are using the EMR dataset extracted from patients' health records of UKY medical center and its affiliated clinics. Our dataset contains information from $\approx$ 922,000 patients between 2004 and the first quarter of 2016. In this dataset, there are 4.15 million patient visits with the average of 6.05 visits for a patient in

a given time interval. We treat each patient's medical information as a transaction where it is a time labeled sequence of patient visits ordered chronologically. Each patient visit contains medications and diagnoses as medical codes, patient demographics and a time label which shows the date when the patient visit occurs. Thus, our dataset has around 922,000 transactions with the longest transaction size of 584 and the shortest transaction size of 1. Therefore, the average transaction size is equal to the average number of visits.

In this study, we are using three data tables from the EMR dataset: medications, diagnoses, and patient demographics. Diagnosis table contains more than 11,770 unique ICD-9-CM codes. These codes are standardized to advance physicians' recordings related to the consistency of the health condition for a patient. Our medication table has 1,397 unique medication codes by *Cerner Multum*^{TM} *Lexicon Plus* codes which are also used by Centers for Disease Control and Prevention for their medical care surveys. Patient demographics table includes beneficial information about the health status of patients. In our study, we are using six important patient demographics for exploring the psychological diseases: gender, BMI score, age, tobacco usage, marital status, and race.

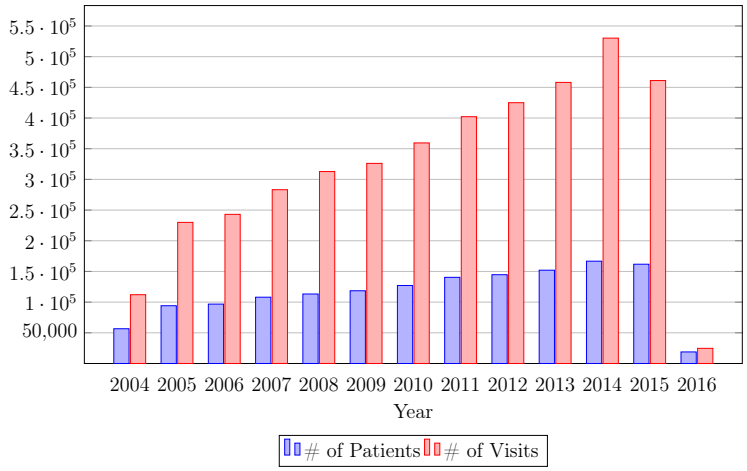


Figure 4.1: Number of visits and patient to the UKY hospital and affiliated clinics for each year

There are total of 15,072 ICD-9-CM codes and our dataset contains 11,770 of these codes. Some diagnosis can be identified with multiple ICD-9-CM codes. Instead of using each code separately, using the group of codes related to the same diagnosis is more beneficial in statistical studies. Although grouping the codes lose precision, it will improve the accuracy and efficiency of the statistical analysis. In order to group ICD-9-CM codes, we use CSS developed as part of the HCUP

Table 4.1: List of ICD-9-CM codes related to targeted disorders

Disorder	ICD-9-CM code
Depressive Disorders	293.83, 296.20, 296.21, 296.22, 296.23, 296.24, 296.25, 296.26, 296.30, 296.31, 296.32, 296.33, 296.34, 296.35, 296.36, 300.4, 311
Anxiety Disorders	293.84, 300.00, 300.01, 300.02, 300.09, 300.10, 300.20, 300.21, 300.22, 300.23, 3002.9, 300.3, 300.5, 3008.9, 300.9, 308.0, 308.1, 308.2, 308.3, 308.4, 308.9, 309.81, 313.0, 313.1, 313.21, 313.22, 313.3, 313.82, 313.83

In our dataset, the medication table contains 1397 unique Multum medication codes. We grouped these codes in higher level categories which yield 190 medication groups. Each variable in the demographics table is grouped according to its value. There are two gender groups: female and male. BMI variable is separated into 4 different groups which are underweight, normal weight, overweight, and obese (Flegal et al., 2014). Age variable categorization ranges for our study are 1-12, 13-17, 18-24, 25-44, and over 44. There are two groups for tobacco usages where patients who have never used tobacco and used tobacco at least once throughout his/her life. Divorced, separated, single, married, and widowed are the categories for marital status. Finally, we have nine race groups. For each variable, we have an unknown category for missing value.

4.3 CAR Mining From Patient Data

In this section, we explain how to discover CARs from the EMR dataset which contains over four million patients data. Our process includes three basic components: (1) generating a list of confounding variables, which are essential for causal discoveries, for each medical item, (2) discovering statistically significant potential CARs using the confounders, (3) analyzing the reliability of our method against domain expert ratings for causality.

4.3.1 Generating Confounders

Confounding variables are independent factors which can affect both the exposure and outcome variables and alter the true relationship between these variables. Age,

gender, weight (BMI), or smoking status are attribtues that are typically considered as the confounders in many health studies.

In order to identify the CAs, extraction of correct confounding variables is crucial. Expert knowledge is the basis for detecting these variables. Beside expert knowledge, there are various algorithms designed to discover the confounders. Low et al. (2016) studied around 20 automated methods to discover confounding variables and the best result was achieved with Lasso Logistic Regression (Tibshirani, 1996), which performs L1 regularization, for an EMR database. L1 regularization includes a penalty based on the sum of the magnitudes (absolute values) of coefficients, which reduces some of the coefficients to zero during training. According to the results, there are two possible dependencies: positive and negative. Positive dependency occurs if the increase in the first (independent) variable causes the increase in the second (dependent) variable. Whereas, negative dependency implies that the increase in the confounder causes the decrease in the medical item or vice versa. The output values, which are correlation coefficients, show the strength of the dependency between a confounder and a medical item. Therefore, we will apply this analysis to outcome variables and each medical code which has a strong association with our outcome variables to find the confounders.

After finding the confounders for each medical item, we order them according to their correlation coefficient to select the strongest common confounders for an exposure and outcome. The process begins with identifying all common confounders from the list of confounders for an exposure and outcome in a way that a confounder has a positive or negative dependency for both items. That is, we keep all items which have positively or negatively dependent on the exposure and the outcome of the rule. Then, we sort the absolute value of results coming from the L1 regularized logistic regression in descending order and keep the rank of each item. In the end, we generate a final score for each confounder by adding the ranking for the exposure and the outcome.

4.3.2 Causal Association Rules

Mining EMRs to identify potentially meaningful and novel clinical associations has been gaining popularity in the medical informatics field. However, it is well known that associations do not necessarily indicate causal relationships. A causal association also indicates that there is a cause-effect relationship between an exposure and outcome of a rule. That is, the change of the exposure causes the variation of the outcome. To obtain CARs, we generate all ARs using our data. Then, we apply

our method (a reranking approach) to the list of ARs. Additionally, we employ the temporal precedence of E with respect to Y. To impose this, we treat a patient's longitudinal record of visits as a transaction.

Let $\mathcal{I}$ be union of all medical items coming from the medications, diagnoses, and patient demographics tables. A set $C = \{i_1, \ldots, i_k\} \subseteq \mathcal{I}$ is called a clinical *item set* with k items and a patient *visit sequence* $S = \{V^1, ..., V^n\}$. Here, $V^j = (vid, C^j, t_j)$ is a patient visit which is defined over $\mathcal{I}$ where *vid* is the patient visit ID, C^j is a clinical item set and t_j is the time of the related patient visit as well as $t_{j-1} < t_j$ which also means V^{j-1} occurs before V^j for values of j greater than one. For our purposes, $T = \{pid, S\}$ is called a *patient transaction* where *pid* is the patient ID and S is a patient visit sequence. Then, we employ temporal restrictions to our patient transactions as follows.

- Logic dictates that the cause has to occur before the effect. Therefore, we remove all medical items which occur at the same visit and visits occurring after the patient has been diagnosed with our target code.
- We apply a washout period of β (specfied usually in months). Thus, we only identify transactions for a patient where the time difference between the first visit and the last visit of the patient is higher than or equal to β. That is, we remove all transactions where a patient who has only one visit to the hospital and does not meet washout period restriction.
- If a patient has never been diagnosed with target condition all medical codes will be used.
- If there is more than one occurrence of the target condition, first diagnosis time will be used.
- We consider items occurring in the same visit with target condition and visits after that as a new transaction. Then, we apply the same temporal restriction to this new transaction. Thus, it is possible for a single patient's visit sequence to result in more than one transaction.

After the temporal restrictions applied, all patient transactions are denoted as the patient database $\mathcal{D}$.

Association Rule Mining

Here, a set E is a clinical *itemset* and S is a patient *visit sequence*. A patient *visit sequence* S is said to contain a clinical *item set* E via the definition

$$contain(E, S) = \begin{cases} \text{true}, & \text{if } E \subseteq (C^1 \cup ... \cup C^k) \\ \text{false}, & \text{otherwise} \end{cases} \tag{4.1}$$

A *patient transaction* $T = (pid, S)$ is said to *support* an item set E if $contain(E, S) = true$ and the *support* of E in a patient database $\mathcal{D}$ is defined as:

$$support(E, \mathcal{D}) = |\{pid : (pid, S) \in \mathcal{D}, contain(E, S)\}|. \tag{4.2}$$

An item set is deemed *frequent* if its support is greater than a given minimum support σ. Thus, the set of frequent item sets with respect to σ is defined as:

$$\mathcal{F}(\mathcal{D}, \sigma) = \{E : support(E, \mathcal{D}) \geq \sigma\}. \tag{4.3}$$

Next, an Association Rule (AR) is a rule of the form $E \Rightarrow Y$ where E and Y are item sets and $E \cap Y = \emptyset$. The *confidence* of an association rule $E \Rightarrow Y$ denoted by

$$conf(E \Rightarrow Y, \mathcal{D}) = \frac{support(E \cup Y)}{support(E)}, \tag{4.4}$$

models the probability $P(Y|E)$ and establishes the association of the outcome item set Y with the exposure item set E. Besides minimum support for item sets, we can establish a minimum confidence γ for ARs and define a stronger notion of frequent and confident ARs. Apart from minimum confidence and minimum support, we implement another measure to select the list of AR over the patient database $\mathcal{D}$ which is odds ratio lower bound (ORLB) with a 95% confidence interval. The set of association rules are

$$\mathcal{R}(\mathcal{D}, \sigma, \gamma) = \{E \Rightarrow Y : E \cup Y \in \mathcal{F}(\mathcal{D}, \sigma), conf(E \Rightarrow Y) \geq \gamma, ORLB_{E \Rightarrow Y} > 1\}, \tag{4.5}$$

which consists of confidence thresholded statistically significant ARs obtained from frequent item sets.

Generating Fair Dataset from EMR database

After selecting a list of the confounding variables, we generate a partly matched dataset called FD (Li et al., 2015a). FD is created by selecting two transactions regardless of the outcome where both transactions have the same values for confounders but different (Boolean) values for the exposure. FD of a rule $E \Rightarrow Y$, where E is the exposure and Y is the outcome, is a sub-database of the EMR database and consists of one to one matched pairs. To obtain pairs, we divide the EMR database into two parts: EMR_E and $EMR_{\neg E}$ without any consideration of the outcome Y. Here, EMR_E consists of exposed transactions and $EMR_{\neg E}$ consists of unexposed transactions. Then, we try to match each transaction in EMR_E with a transaction in $EMR_{\neg E}$ where both transactions have the exact same values for confounding variables. That is, we are not using any similarity measures. During the matching process, we randomly pick a transaction from EMR_E and try to identify a matched transaction in $EMR_{\neg E}$. Finally, we move all such matched transaction pairs to the FD.

Finding Causal Rules

Based on the generated FD, we create 2×2 contingency table for the rule $E \Rightarrow Y$. Here, our rows will represent transactions from EMR_E database and columns represent the transactions from the $EMR_{\neg E}$ database. Each transaction has two outcome values as shown in Table 4.2: Y indicates a true outcome and $\neg Y$ expresses a false outcome. In this table, each count corresponds to a pair. Therefore, the size of FD is calculated as

$$|FD| = 2 \times n \tag{4.6}$$

where n is the number of total matched pairs.

Table 4.2: 2×2 Contingency table for a rule $E \Rightarrow Y$ on FD

		$EMR_{\neg E}$		
		Y	$\neg Y$	Total
EMR_E	Y	a	b	a+b
	$\neg Y$	c	d	c+d
	Total	a+c	b+d	n

To create the contingency table, there are four possible values for each pair $P = (T_1, T_2)$ where transaction $T_1 \in EMR_E$ and transaction $T_2 \in EMR_{\neg E}$. The value of "a" in Table 4.2 corresponds to the number of pairs where both transactions have the true outcome value while the value of "b" is the number of pairs where T_1 has the true outcome value but T_2 has the false outcome value. The value of "c" corresponds to the number of pairs where T_1 has the true outcome value and T_2 has the false outcome value while the value of "d" is the number of pairs where both transactions have the false outcome value.

After building the contingency table, we will calculate the Odds Ratio (OR) score to determine whether a rule is potentially causal. OR is commonly used to measure the association strength of the exposure, a disease or medication, with the outcome, a medical condition (Morris and Gardner, 1988). Table 4.2 shows a 2×2 contingency table for a CAR. Using the values of the contingency table, calculating OR score gives the causal association level between an exposure and an outcome. OR of a CAR (Fleiss et al., 2013) is calculated as

$$OR_{FD} = \frac{b}{c} \tag{4.7}$$

where $E \Rightarrow Y$ is a CAR candidate while b and c are the values appear in Table 4.2. An association between the exposure and the outcome is identified according to the calculated OR value. There are three possible values of the calculation. If $OR = 1$, there is no association between the exposure and the outcome because odds of the outcome for the exposure is equal to odds of the outcome for not exposure. $OR > 1$ indicates that there is a positive association between the exposure and the outcome as well as the odds of the outcome significantly increases when it is exposed. When $OR < 1$, there is a negative association between variables and the odds of the outcome is lower for the exposure.

Evaluation of a rule by OR alone is not always enough to ascertain the association between the exposure and the outcome especially for small FDs. Therefore, we also calculate 95% confidence intervals to estimate the variation of the OR. The formula to calculate standard error (SE) of 95% confidence intervals for ln(OR) is defined as

$$SE_{FD}(\ln(OR_{FD})) = \sqrt{\frac{1}{b} + \frac{1}{c}} \tag{4.8}$$

Using SE_{FD} value, the OR lower bound (ORLB_{FD}) and OR upper bound (ORUB_{FD})

for FD of the 95% confidence interval of OR are

$$ORLB_{FD}(E \Rightarrow Y) = e^{\left(\ln(OR_{FD}) - 1.96 \times SE(\ln(OR))\right)}. \tag{4.9}$$

and

$$ORUB_{FD}(E \Rightarrow Y) = e^{\left(\ln(OR_{FD}) + 1.96 \times SE(\ln(OR))\right)} \tag{4.10}$$

For a rule to be considered causal, $ORLB_{FD}$ must be greater than 1. The rationale for using ORLB over OR is that ORLB balances the assurance that the result is not due to the chance with the strength of the estimated effect considering the variance of the estimator. Another advantage is to avoid misjudgment caused by very small b and c values. As seen on equations 4.8 and 4.9, using SE and 95% confidence intervals gives a penalty equivalent to the square of the magnitude of the total multiplicative inverse of b and c to $\ln(OR)$. That is, lower values will increase the penalty score.

4.4 Experiments and Results (Quantitative & Qualitative Analysis of CARs)

In our experiments, we used the EMR dataset which has $\approx$ 922,000 patient transaction data collected between 2004 and 2016. Before running experiments, we applied 6 months of temporal restriction ($\beta = 6$) to choose patients with known enough medical history to employ our method. That is, we only use the patients with at least six-month medical history in our EMR database. When we remove all patients with the time between the first and last visit is less than 6 months, people with only one visit to the hospital are also be removed automatically. Consequently, we avoid the bias that may occur because of patients with insufficient medical history. Then we apply our method to mine all causal rules from the EMR database.

To generate CARs, we obtain all ARs with the support threshold $\sigma = 100$, confidence threshold $\gamma = 10\%$ and ORLB > 1. In other words, we extract all rules with at least 100 occurrences in all transactions where at least 10% of transactions that contain exposure also contain outcome, and are statistically significant based on 95% confidence intervals. When AR mining approaches are applied to clinical datasets, these configuration settings yield optimum performance as in earlier studies (Wright et al., 2010; Wright et al., 2013). The ARs are mined by the Linear-time Closed item set Miner (LCM Ver. 3) by Uno et al. (2005) that utilizes a combination of convenient data structures. Then, the OR restriction is applied.

We conduct experiments for two outcome variables: depressive disorders and anxi-

ety disorders. We chose these outcomes are well known to be causing highest disability among mental conditions and are also well represented in our dataset. Additionally, our domain expert raters suggested these mental conditions for our research based on their experiences as practicing psychiatrists. We chose to use our exposures and outcomes as singletons. Besides, very long rules are not interesting as they capture highly specific scenarios that are not amenable to reasonable interpretation and typically have low support as noted in prior efforts (Hämäläinen, 2011).

In our effort, statistical strength is measured by OR or more specifically ORLB. A rule called statistically significant if the ORLB value is greater than 1. As we mentioned in section 4.3.2 to estimate the precision of the OR, calculating 95% confidence intervals is suitable for observational studies.

4.4.1 Causality Scores

As we discussed in section 4.3.2, we first generate all confounding variables for each possible exposure and two outcomes which are depressive disorders and anxiety disorders. Then, we create the corresponding FD to compute the causality score. To generate FDs, we use 11 different confounder sets presented in Table 4.3 based on ranking of informativeness of confounders. In this table, c_1 is the strongest common confounder between an exposure and outcome as well as c_2 is the second strongest common confounder between related exposure and outcome and so on. To decide the strongest common confounders, we remove all confounders from the list for exposure and outcomes which do not occur in both lists. Subsequently, remaining items are listed in descending order of the absolute value of coefficients calculated by lasso regression. Then, we apply a rank quality metric, discounted cumulative gain (DCG) (Järvelin and Kekäläinen, 2002, Sections 2.2–2.3) to each confounder list. The final order of confounders is the total DCG values of exposure and outcome in descending order. Later, we calculate ORLB value of the rule for each confounder set to calculate final causality score — the average of these 11 ORLB values.

As shown in Table 4.4, we identified 300 and 313 statistically significant ARs for depressive disorders and anxiety disorders respectively. From these ARs, there are 61 CARs for depressive disorders and 39 CARs for anxiety disorders. In our EMR dataset, we have 216 ARs where the exposure is a diagnosis code and the rest have a medication code as an exposure for depressive disorders. For anxiety disorders, we have 228 ARs having a diagnosis code as an exposure and 85 ARs where the exposure is a medication code. We identify 57 CARs and 39 CARs having a diagnosis code as an exposure for depressive disorders and anxiety disorders, respectively. The number

Table 4.3: The list of criteria to create FD

Method	Confounders
1	Demographics
2	$c_1 \cup$ Demographics
3	$c_1, c_2 \cup$ Demographics
4	$c_1, c_2, c_3 \cup$ Demographics
5	$c_1, ..., c_4 \cup$ Demographics
6	$c_1, ..., c_5 \cup$ Demographics
7	$c_1, ..., c_6 \cup$ Demographics
8	$c_1, ..., c_7 \cup$ Demographics
9	$c_1, ..., c_8 \cup$ Demographics
10	$c_1, ..., c_9 \cup$ Demographics
11	$c_1, ..., c_{10} \cup$ Demographics

of CARs where the exposure is a medication code is 4 for depressive disorders and there is no medication code as an exposure for anxiety disorders. Finally, the average ORLB score of CARs is 1.368 for depressive disorders and 1.3 for anxiety disorders.

Table 4.4: Statistics for CARs

	Depressive Disorders	Anxiety Disorders
# of ARs	300	313
# of CARs	61	39
# of diagnosis in ARs	216	228
# of medication in ARs	84	85
# of diagnosis in CARs	57	39
# of medication in CARs	4	0
Average ORLB of CARs	1.367513	1.300033

4.4.2 Domain Expert Assigned Plausibility Scores

To validate our final causality scores, we collaborated with two practicing psychiatrists. Each assigned a graded ordinal score of 1 through 5 where a high score indicates a causal relationship from the perspective of either phenotypical or bimolecular plausibility. The interpretation of the graded score follows:

1. No evidence or possible explanation of causality,
2. Very minor evidence of plausible causal nature,
3. Fair evidence with some plausible mechanism of causality,

4. Moderate evidence of a trail(s) leading from the candidate to outcome,

5. Clear (well known or commonly accepted) evidence of causal pathways between the candidate and depression.

Regarding human expert evaluation, we are essentially looking at informed assessment based on practice, research, literature exposure, and medical education for a potential causal connection between each candidate and depressive disorders or anxiety disorders. We gave them the exposure list of all statistically strong ARs for both disorders without knowing the strength of neither the association nor the CA. The list is ordered alphabetically not to leak any hints to the raters about the associations. Also, our experts scored the rules without communicating with each other since we did not want raters influencing each other.

According to the graded scores we collected from our raters, the score of 5 did not occur for Anxiety while it occurred only once for depressive disorders which is given by R_1 to bipolar disorders. This outcome indicates that according to the knowledge of R_1 there is a clear evidence of causal pathways between bipolar disorders and depressive disorders. There are 9 and 8 codes each with the score difference of two between the raters for anxiety disorders and depressive disorders respectively. The scores assigned by each expert is listed in Table 4.5. There are only three codes scoring a 4 by both raters for depressive disorders. From the list of exposures, 110 codes have the common score of 1 for depressive disorders and 168 codes have the common score of 1 for anxiety disorders.

Table 4.5: Scores assigned by raters for CARs

	Rater 1					Rater 2				
Score	1	2	3	4	5	1	2	3	4	5
Depressive Disorders	168	93	27	11	1	129	124	38	9	0
Anxiety Disorders	215	89	15	4	0	213	70	28	2	0

Typically IRR scores are calculated to measure the reliability of raters if more than one rater is used. In this study, we calculated three well-known and reliable measures: Cohen's Kappa, Spearman's Rho, and Gwet's AC1. Table 4.6 shows the result for five measures and Table 2.5 shows the agreement level for IRR measures. Since our rater scored each possible exposure with the scale of 1 to 5, weighted measurements will give more reliable scores. Both probabilistic measures, standard κ and Gwet's AC1, give the same penalty to all disagreements to calculate the IRR score. However,

weighted scores give appropriate weights to the disagreements according to the score differences.

Table 4.6: IRR scores for raters

	Depressive Disorders	Anxiety Disorders
Cohen's Kappa(κ)	0.34795	0.30366
Weighted κ	0.63249	0.51550
Gwet's AC1	0.5239	0.61463
Weighted Gwet's AC1	0.92924	0.95131
Spearman's Rho (ρ)	0.61415	0.46689

These results indicate that according to the Gwet's AC1 score, our raters have almost perfect agreement for depressive disorders and anxiety disorders with values 0.93 and 0.95 respectively. Weighted scores reveal better results for both measurements and outcomes. Only standard κ shows that the agreement level between raters is fair. While weighted κ and ρ indicate substantial agreement level for depressive disorders, Gwet's AC1 reveals a moderate agreement between raters. For anxiety disorders, weighted κ and ρ indicate that there is a moderate agreement between raters, whereas Gwet's AC1 shows a substantial agreement.

4.4.3 Comparison of Scores

Our goal is to find medical codes which have a CA between an exposure and our target code using only the EMR dataset. Essentially, we compared our causality scores with graded scores by domain experts to test the accuracy of our method.

Table 4.7: Scores assigned by raters for CARs

	Rater 1					Rater 2				
Score	1	2	3	4	5	1	2	3	4	5
Depressive Disorders	22	24	12	3	0	13	30	16	2	0
Anxiety Disorders	15	16	5	3	0	13	13	12	1	0

From the list of exposures, 110 of them have the score of 1 from both raters for depressive disorders and only 9 of them appeared in our list of CARs. 168 possible CARs have the common score of 1 for anxiety disorders and 10 of them are found causal according to our analysis. As shown in Table 4.5, our method finds 3 of 4 exposures with a score of 4 for anxiety disorders. Table 4.7 reveals the information about our expert scores for CARs. It also indicates that a high number of ARs with

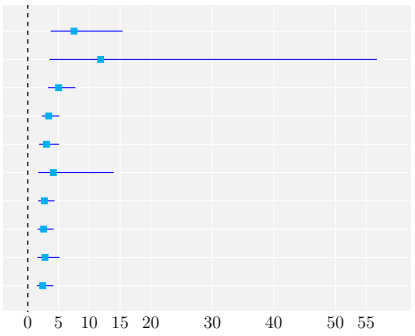
low scores is not found causal. On the other hand, anxiety disorders which scored a 4 by both raters have the third highest ORLB value and are considered as causal for depressive disorders. Whereas bipolar disorders which are the only exposure we have with the score of 5 did not appear in our CAR list.

Table 4.8: Rater graded scores

	Depressive Disorders	Anxiety Disorders
Avg. score R_1	1.61333	1.41853
Avg. score R_2	1.75667	1.42173
Avg. of Averages	1.685	1.42013
Avg. score for causal values R_1	1.93443	1.89744
Avg. score for non causal R_1	1.53333	1.35273
Avg. score for causal values R_2	2.11475	2.02564
Avg. score for non causal R_2	1.67083	1.33455

Table 4.8 shows the average values of expert graded ordinal ratings for potential CARs for both depressive disorders and anxiety disorders. The average score of rater 1 (R_1) is 1.61333 for depressive disorders and 1.41853 for anxiety disorders. The average score is 1.75667 and 1.42173 for depressive disorders and anxiety disorders respectively by rater 2 (R_2). The average score of both raters is 1.685 for depressive disorders and 1.42013 for anxiety disorders. Moreover, the average graded score of casual rules, which is discovered by our method, and non-causal rules are also shown in Table 4.8. These scores show that causal rules are scored 40% and 51% higher for anxiety disorders and 26% and 27% higher for depressive disorders by R_1 and R_2 respectively.

Exposure	OR	ORLB	ORUB	Avg. Score	OR and 95%CI
ADD and ADHD	7.5175	3.7254	11.3096	2.5	
Suicide and intentional self-inflicted injury	11.8428	3.5085	20.1771	2	
Anxiety disorders	5.0126	3.2774	6.7478	4	
HIV infection	3.4097	2.2729	4.5465	2.5	
Leukemias	3.0407	1.8187	4.2627	1.5	
Sleep disorders	4.1495	1.6748	6.6242	3	
Spondylosis; IDD; other back problems	2.7011	1.6717	3.7305	1	
Schizophrenia and other psychotic disorders	2.5838	1.6018	3.5658	2.5	
Esophageal disorders	2.801	1.5605	4.0415	1	
Antidepressants	2.4287	1.5066	3.3508	2	



Abbreviations:ADD,Attention deficit disorder. ADHD,Attention deficit hyperactivity disorder. IDD,Intervertebral disc disorders.

Figure 4.2: Forest plot of top 10 exposures for depressive disorders

Exposure	OR	ORLB	ORUB	Avg. Score	OR and 95%CI
Depressive disorders	3.9888	2.1324	5.8452	3	
Intellectual disabilities	5.6357	1.9468	9.3246	2.5	
Somatoform disorders	6.9415	1.8418	12.0412	2.5	
Bipolar disorders	3.7351	1.7889	5.6813	2.5	
Other congenital anomalies	11.3467	1.6715	21.0219	1.5	
HIV infection	4.1771	1.6271	6.7271	2	
Substance-related disorders	3.2991	1.5835	5.0147	2.5	
Communication disorders	5.1623	1.5812	8.7434	2.5	
Other infections; including parasitic	3.4891	1.5811	5.3971	1	
Anal and rectal conditions	5.5109	1.568	9.4538	1	

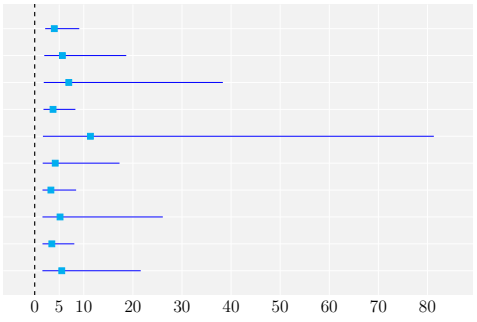
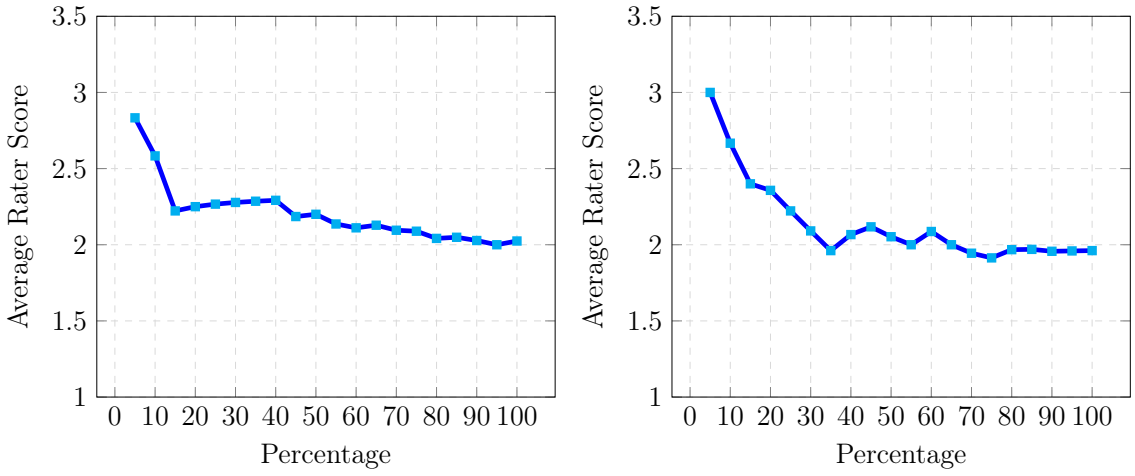


Figure 4.3: Forest plot of top 10 exposures for anxiety disorders

In Figures 4.2 and 4.3, we show forest plots of top 10 exposures for depressive disorders and anxiety disorders respectively. Squares indicate the OR values and the lines present the error range. In these figures, exposures are ordered by the ORLB values. That is, exposures with high OR do not necessarily rank higher due to the confidence interval based ranking. We also compute the average expert scores for CARs considering the top 5%, 10%, ..., 100% of the CARs cumulatively with 5% increases for both depressive disorders and anxiety disorders. Figure 4.4a shows the result for depressive disorders and figure 4.4b shows the result of anxiety disorders. Clearly, when we consider all causal rules the average is 2.02 and 1.96 for depressive disorders and anxiety disorders respectively. Figures 4.4a and 4.4b indicates top CARs tend to have higher average scores from our experts.

4.5 Conclusion

In this chapter, we presented the CAR mining approach to extract all rules from the EMR database, which contains data from more than 900,000 patients with visits between 2004 and 2016, for anxiety disorders and depressive disorders as outcomes. We exploited an AR mining technique to generate all statistically significant ARs as part of our possible CAR universe. We built multiple confounder lists each of which consists of all 6 different demographic and life style variables (gender, BMI score, age, tobacco usage, marital status, and race) and related top common medical codes ordered by the lasso regression score presented in Table 4.3. Then, we generated the FD which consists of matched transaction pairs for each confounder criterion to calculate the causality score — the ORLB of the 95% confidence interval, from the FD. We collaborated with two practicing psychiatrists to rate all ARs without any



(a) Depressive disorders

(b) Anxiety disorders

Figure 4.4: Average score for related percentage for anxiety disorders and depressive disorders

statistical information. They graded 300 rules for depressive disorders and 314 rules for anxiety disorders. Our scores indicate that 62 of 300 rules are causal for depressive disorders while 39 of 314 rules are causal for anxiety disorders. We presented two forest graphs in Figures 4.2 and 4.3 to show the average causality scores of the best 10 medical codes. Our results demonstrate that medical codes with a higher causality score tend to have a higher plausibility rating from medical experts, thus demonstrating that computational methods can be used to identify potential causal associations from EMRs.

Chapter 5 Predictive Modeling through Sequential Patterns and Recurrent Neural Network (RNNs)

Being able to accurately measure the risk of a future incidence of a disease before it fully manifests is arguably one of the most important applications of machine learning for healthcare. In general, most people schedule regular visits to a healthcare provider to be able to spot early symptoms of a condition. However, predicting a patient's future diagnosis is an arduous task for a busy physician due to the limited resources during the short visit. Although doctors have useful information about a patient including lab results, previous conditions, social changes, demographics, considering all the data about the patient before making a decision is sometimes not feasible. Hence, an automated early detection system of a disease, which considers as much patient data as possible before making a prediction is of immense utility to assist physicians in their patient interactions.

Prediction is at the core of machine learning and there is a rich history of combining classical data mining approaches such as ARM with machine learning under the umbrella of *associative classification* (Yin and Han, 2003; Yang et al., 2016; Letham et al., 2013). Deep learning methods (Choi et al., 2015; Choi et al., 2016b; Choi et al., 2016c; Che et al., 2016) are also being independently studied for predictive modeling in biomedicine. Here, we combine sequential pattern mining (Ghosh et al., 2016; Wright et al., 2015) with RNNs to predict a future diagnosis of chronic conditions (with depressive disorders use-cases). The central idea is to model a patient's medical history as a chronologically ordered sequence of transactions each of which corresponds to a visit and consists of clinical variables (diagnoses, medications, procedures, and visit demographics). Additional gap constraints are also imposed to ensure that any two consecutive item sets in the sequence are not too far apart (temporally) in terms of absolute time difference. Based on an experimentally chosen minimum support, a few top sequential patterns are chosen and then used as features in machine learned models. However, we also imagine these sequential patterns forming a meta-sequence by themselves based on the windows in time spanned by each pattern. This makes it amenable to use recurrent neural networks that compose information available in items that form a sequence. In order to assure that we can predict a target condition sufficiently ahead of time, we will restrict the prediction window to 12 months before the actual diagnosis date. A variant of sequential patterns, so called *contrast* sequential patterns are used; contrast patterns are inherently expected to

be more prominently occurring in the positive instances than the negative instances. Contrast pattern mining also facilitates a new way of going about disease prognosis. By splitting a patient's chronological record into two halves, we can create two databases corresponding to subsets where the target condition is present and absent; the first set is the positive database to find contrast sequential patterns. By focusing on patterns that manifest more prominently with the diagnosis of a condition, we can analyze different phenotypes of a condition and the corresponding courses as the condition is treated.

In this effort, our specific contributions are:

- In current literature, much attention has not been paid to the effects of minimum history available, inter-visit gaps, washout periods, and prediction horizons. Using various settings for these variables, we create multiple subsets of our EMR dataset with depressive disorders as the outcome.
- We apply different RNN based models along with V-LSTM model (from Section 2.7.1) to assess their relative performances in each of the configurations identified earlier.
- We also experiment with different ways to embed the input EMR code sets and the unchangeable demographic variables (e.g., gender, race).
- We develop a novel method that leverages sequential contrast patterns (SCPs) to build hierarchical RNNs that compose SCPs and then a sequence of chronologically ordered SCPs, encapsulating each visit.
- We demonstrate that a novel architecture that combines the SCP based model and the V-LSTM model produces the best predictive performance.

5.1 Related works

Several efforts apply SPM and SCP approaches for either classification or prediction purposes. Cheng et al. (2016) applied SPM approach to early detection of COPD using National Health Insurance Research Database of Taiwan. This database contains over 0.9 million patients' medical history records coded by ICD-9 where each patient visit consists of maximum of three ICD-9 codes. They implemented an SPM algorithm called SPADE (Ayres et al., 2002) to mine all COPD related sequential patterns to classify a new instance as COPD or non-COPD. Then, they verified their COPD related rules by searching PubMed article count for each rule and used this count as

their novelty assessment. Hanauer and Ramakrishnan (Hanauer and Ramakrishnan, 2013) mined temporal relationships from the EMR database using diagnosis history of patients coded by ICD-9. These temporal relationships include two different approaches. First, they found one to one associations applying item wise and pair wise minimum support thresholds and then they chose p-value, χ^2, and OR as statistical measures of significance. Second, they applied temporal analysis to each pair with five different time ranges between the diagnosis of antecedent and consequent. These time ranges are between 1 day and 10 years. Ghosh et al. (2016) applied sequential contrast mining approach to predict hypotension risk of Intensive Care Unit patients. They divided their dataset into two parts: positive sequences and negative sequences. Afterwards, they identified all SPs from each part and compared these patterns to generate all SCPs which are used for classification. Wright et al. (2015) apply SPM techniques to a medication database to predict the next prescribed medication for the patients using their history of medications. They used diabetes and medications related to diabetes as a test case due to the progressive nature of such disease. A step wise pharmacological management technique is applied to control and avoid exacerbation of the disease. In this research, they studied both generic medication codes and class level medication codes separately. They generated SPs and ordered them according to the frequency. For a patient, they output best possible predictions according to patient's medication history. The class level approach gave almost 50% better result than the generic approach. Reps et al. (2012) applied SPM approaches to an EMR database in order to predict future illness of a patient. Besides the diagnosis of each patient, they also considered a couple of demographic attributes, gender and age.

SCP mining process consists of multiple steps and the most time consuming part is counting sequences to determine the support value. Due to the size of an EMR database, even a typically less expensive task like FIM can take multiple hours. Current approaches, which are computationally expensive and time consuming, for mining SCPs on CPU fail due to the large size of an EMR database or the type of EMR data structure. To overcome this struggle, researchers adapted FIM approaches to use GPU for the frequency counting part (Teodoro et al., 2010; Zhang et al., 2011). They conclude that the same task can be completed up to 173 times faster than when using CPU implementation. Although FIM and SCPM approaches are different, they share some similarities in their formulation. The database used in these approaches can be represented as bit vectors and the support can be calculated by logical operations. In order to increase the processing speed, we adopt minimal distinguishing sequential

pattern mining algorithm (Ji et al., 2007) that runs on both CPUs and GPUs.

Choi et al. (2016c) used a GRU version of RNN to predict health failure using patients' medical records. They use the medication, diagnosis, and procedure history of a patient as input to their method and predict the heart failure as a binary outcome. Their results show that deep learning approaches are better at predicting future health conditions than other well known techniques such as MLP, SPM, KNN, and LR. Choi et al. (2015) designed a multi-label classification model using GRU to predict future diagnosis, medication, and procedures of a patient as well as predicting a patient's next visit time to the hospital. They evaluated their measure by recall@k for k values of 10, 20, and 30 where top-k recall is equal to the number of true positives in the top-k predictions divided by the total number of true positives. Choi et al. (2016b) implemented an algorithm adopting two-level attention mechanism in reversed time order called RETAIN that focuses on visit level and code level signals. They implemented both attention mechanisms using an RNN structure and they combined all three values: visit level attention, code level attention, and the RNN output to make the final prediction. A graph based attention mechanism was implemented for healthcare representation learning in Choi et al. (2016a). In this study, the embedded matrix was generated using the graph-based attention mechanism, and then the input is transformed by the embedded matrix before being fed to the GRU unit. Lipton et al. (2015) employ a LSTM model for the multilabel classification task for predicting ICD-9 codes using patients' clinical time series data. They used these continuous data to predict the relationship between health episodes and diagnosis codes. An EMR database contains both static and dynamic information about the patients. All of the research studies using a RNN model discussed the need to focus on dynamic information and missing static information. Esteban et al. (2016) created a prediction mechanism by combining an MLP model and a GRU model together to predict a future event. They input patients' static information into an MLP model and dynamic information into a GRU model for each type of data. DeepCare (Pham et al., 2017) is another recent model to predict future diagnosis of a patient with concentration of the patient's previous conditions, medications, procedures, and the time between two consecutive hospital visits. They cleverly transformed LSTM structure to the C-LSTM to handle each patient visit data in a time point. In this study, there are two cohort studies designed for the EMR database: mental health and diabetes. They compare their results with a Markov model, an LSTM, and a regular RNN and show improvements over these models.

5.2 The EMR Database and Cohort Selection

In this study, we obtain the EMR data from the UKY medical center and its affiliated clinics which contains 14.3M patient visits for 1.12M patients between 2004-2018. The average visit number for each patient is 12.75. We used three tables from the EMR database and the demographics information of the patients. First Table is the ICD-9 and ICD-10 coded diagnosis codes grouped by the CSS software. Second table has medication codes specifically *Cerner Multum*^{TM} *Lexicon Plus* codes in which we used the hierarchical structure to group. Third table has procedure codes specifically CPT codes which are also grouped by CSS software. We also employ three different visit level demographics of patients: tobacco usage, age, and BMI. Tobacco usage of a patient is grouped into 3 groups: yes, no and unknown at that visit. Table 5.2 shows six different age groups we used in this study and ten different BMI groups.

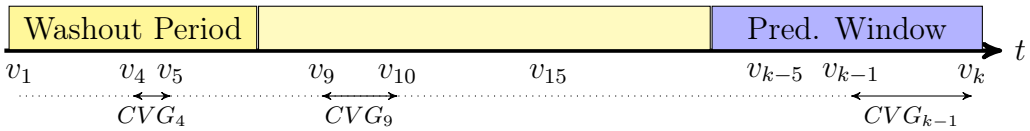


Figure 5.1: Patient visit history

Table 5.1: List of variables and possible values to generate database variations.N/S used as value when there is no limitation specified.

Variable	Posible Values
Min. number of visits	N/S, 10, 20, 30
Max. gap	N/S, 3, 6, 12
Observation window	6, 12
Washout Period	6

First part of this study showed us that longer patient records give a better prediction accuracy. Therefore, we decided to crate a cohort from this EMR database with respect to four variables. Figure 5.1 shows how variables are measured from a patient's visits. First variable is the prediction window which is the time horizon (in number of months) into the future when we would like to make a prediction. This is a period of time between last patient visit or the occurrence of the target disease in patient's medical history and the last visit our model is using to predict. In our work, we are focusing on 12 months for the time horizon (prediction window). That is we want to predict a patient's future diagnosis a year before it actually occurs. Second

variable is the washout period which is the minimum number of months a patient must be in our system to be included in our input and when curating training data this also implies the minimum time before the first diagnosis of the target condition is made as per the EMR data. The third variable is the maximum consecutive visit gap (CVG) which shows the time difference between consecutive patient visits to the hospital. That is $\forall\ j = 1, \ldots, (l-1)$, we require $t_{j+1} - t_j < k$ for some small k (months) for a patient with l visits. With maximum CVG of 12 months, we choose only patients whose consecutive visits are at most CVG months apart. The last variable is the minimum number of visits a patient has in our system. We chose this variable as 30 to remove patients without enough number of total visits for us to accumulate enough clinical history.

Table 5.2: Classes for BMI and age of a patient

Class Type	Classes
BMI	($<$15), ($15 \leq \cdots < 16$), ($16 \leq \cdots < 18.5$), ($18.5 \leq \cdots < 25$), ($25 \leq \cdots < 30$),($30 \leq \cdots < 35$),($35 \leq \cdots < 40$), ($40 \leq \cdots < 55$),(unknown)
Age	(0-12),(13-17),(18-24),(25-44),(45 and over)

After applying these limitations 26,705 total patients are part of the final dataset. Among these patients 5,405 of them are diagnosed with DD which form our positive cases and remaining 21,300 of them constitute the negative instances of the study. Including the settings described earlier, we created 32 different settings of variables to generate 32 different cohorts using all combinations of variable values given in Table 5.1. Total number of patients ranges from 1,729 and 412,454 depending on the choice of parameter settings. Table 5.3 shows the total number of patients, positive class and negative class sizes for each combination of possible values of variables.

To measure the performance of our models, we split each dataset into four parts: training, validation$_1$, validation$_2$, and test with the corresponding proportions as 65%:10%:10%:15%. In each epoch, we use training data to train the model. Then, validation$_1$ data is used to decide best threshold value before prediction and finally validation$_2$ data with the threshold is used to generate the final score for this epoch. After going through all epochs, we selected the training cut off and model corresponding to the best score on validation$_2$ data. This model is then applied to the test data which has never been used in any stage of the model building process.

Table 5.3: Positive and negative class sizes for each cohort

Min. Visit	CVG	Pred.Window	Washout	Total	Positive	Negative
0	0	6	6	412454	34013	378441
10	0	6	6	236696	27241	209455
20	0	6	6	137149	21309	115840
30	0	6	6	89613	17096	72517
0	3	6	6	4165	643	3522
10	3	6	6	4135	643	3492
20	3	6	6	3479	635	2844
30	3	6	6	2775	602	2173
0	6	6	6	23765	2919	20846
10	6	6	6	21478	2890	18588
20	6	6	6	14635	2724	11911
30	6	6	6	10331	2467	7864
0	12	6	6	81717	8493	73224
10	12	6	6	63831	8113	55718
20	12	6	6	41133	7288	33845
30	12	6	6	29181	6451	22730
0	0	12	6	369757	30935	338822
10	0	12	6	220998	24930	196068
20	0	12	6	131795	19677	112118
30	0	12	6	86966	15901	71065
0	3	12	6	2042	350	1692
10	3	12	6	2041	350	1691
20	3	12	6	1939	349	1590
30	3	12	6	1729	343	1386
0	6	12	6	13896	1941	11955
10	6	12	6	13510	1937	11573
20	6	12	6	10936	1889	9047
30	6	12	6	8408	1799	6609
0	12	12	6	56368	6383	49985
10	12	12	6	49836	6278	43558
20	12	12	6	36124	5898	30226
30	12	12	6	26705	5405	21300

5.3 Methods

In this section, we explain the process of creating SP based database after generating SCPs from EMR database, building the two-level hierarchical LSTM model which uses that database as input beside V-LSTM, and finally combining these two models to increase predictive performance.

5.3.1 Support Counting on GPUs and Parallel Reduction

Counting the support of a candidate pattern is the most time consuming part of the SPM process. We adapt this part to run on a GPU to accelerate the process given GPUs can handle multiple comparisons running in parallel.

We altered the ConSGapMiner algorithm (Ji et al., 2007), which heavily depends on bitwise operators, to mine our EMR database. This algorithm only uses single item at each time point. However, our EMR database has one or more items at each time point as a visit typically leads to multiple codes. Therefore, we need

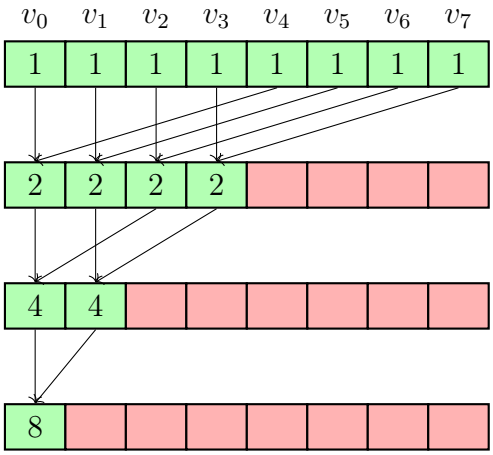


Figure 5.2: Unfolding parallel reduction steps

to implement I-extension as in (Ayres et al., 2002) beside the S-extension which is currently implemented. The process other than support counting follows the same sequence with the original application in parallel because until that point there is a concern of memory conflicts. On the other hand, support counting in a parallel manner requires more attention. To run support counting parallelly, we used the parallel reduction technique which is also useful to find minimum or maximum element in a large dataset in parallel. We can derive from the Figure 5.2, time complexity of support counting using parallel reduction is $O(\log_2 n)$ while counting sequentially is $O(n)$.

5.3.2 Creating Sequential Pattern Based Database

We will transform each longitudinal transaction T to an SCP based transaction, T_{SCP}, where each item of T_{SCP} will be an SCP instead of medical codes and a time for each visit. After finding the best n patterns, we will go through each transaction and convert it to a sequence transaction. Beginning with the first item, we will add each SCP in order of its appearance in the transaction, as shown in Figure 5.3. Our new transaction will be like the one shown in Figure 5.4. So our SCP based transaction is a sequence of sequences.

In order to convert each transaction T to an SCP based transaction, we will generate a set of rules to compare two SPs with each other and decide which sequence will come before another one. Then, we will generate ordered sequences using this set of rules. This comparison involves 13 possible scenarios studied in (Nebel and Bürckert, 1995). To generate ordered sequences, we can summarize these 13 scenarios into seven possible occurrences which are shown in Table 5.4. That is, we will not use

Sequence Relationship	Start&End Points of Sequences	First Sequence
SP_1 equal SP_2	$SP_1^- = SP_2^- < SP_1^+ = SP_2^+$	Choose Random
SP_1 before SP_2	$SP_1^- < SP_1^+ < SP_2^- < SP_2^+$	SP_1
SP_1 meets SP_2	$SP_1^- < SP_1^+ = SP_2^- < SP_2^+$	SP_1
SP_1 overlaps SP_2	$SP_1^- < SP_2^- < SP_1^+ < SP_2^+$	SP_1
SP_1 during SP_2	$SP_2^- < SP_1^- < SP_1^+ < SP_2^+$	SP_2
SP_1 starts SP_2	$SP_1^- = SP_2^- < SP_1^+ < SP_2^+$	SP_1
SP_1 finishes SP_2	$SP_2^- < SP_1^- < SP_1^+ = SP_2^+$	SP_2

Table 5.4: Sequence ordering criteria

both SP_1 *before* SP_2 and SP_2 *after* SP_1 because the *before* condition will be sufficient to decide which sequence will come first and we will omit the *after* condition. For a SP, let SP^- is the starting point and SP^+ is the ending point of that sequence. We will use "$SP_1^- = SP_2^-$" to indicate that both sequences start at the same time, and "$SP_1^- < SP_2^-$" to indicate that SP_1 starts before SP_2. In general, we will be looking for the starting point of each sequence and choose the one that starts first. If both sequences start at the same time, we look for the end points and choose the one that finishes first. If both sequences start and finish at the same time, we choose the ordering randomly.

In this structure, each SCP is an SP and, this gives us the opportunity to implement a hierarchical RNN structure. We convert each item or item groups in SCP to a multi-hot vector where the corresponding bits for an item will be 1 if it's present and the remaining elements will be 0. This multi-hot vector will be the input for the first level RNN and the intermediate output will be the fixed size n vector for the corresponding SCP. Then, we will use this new vector as an input to the second level RNN that composes the SCP representations.

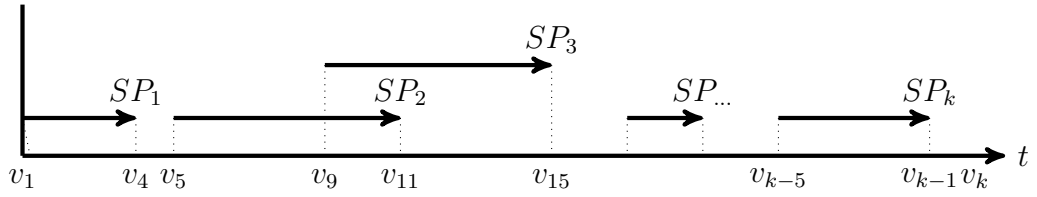


Figure 5.3: Sequence of sequential patterns encapsulating a patient's visit history

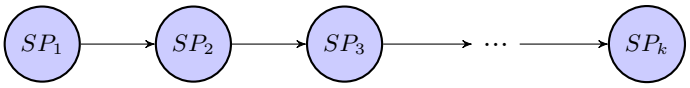


Figure 5.4: Example ordered T_{SCP} based on patterns in Figure 5.3

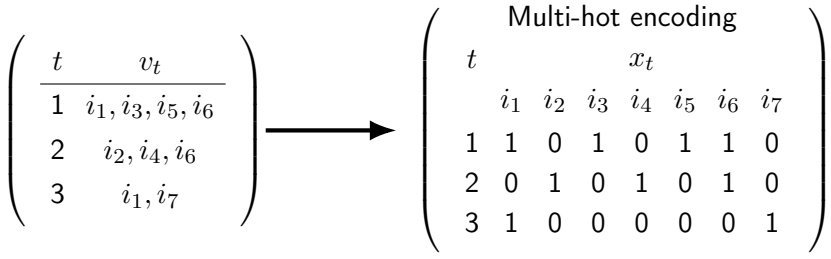


Figure 5.5: Transformation to the multi-hot representation

5.3.3 Input Representations

The raw EMR data goes through a pre-processing step to be converted to the desired form to be fed to the neural models. A few input representation methods maybe be designed for the EMR data. In this part, we describe two previously used representations and a new representation we create for this study. Let $\mathcal{I}$ be the union of all medical codes that can be used for patients. For our purposes, v_j is a set of items and t_j is the time point of the visit then a pair w_j consists of a set of items and visit order index $w_j = \{v_j, t_j\}$ or $w_j = \{(i_1, i_2, ..., i_l), t_j\}$ where $v_j = \{i_1, i_2, ..., i_l\} \subseteq \mathcal{I}$ for $l > 0$ and a set $W = \{w_1, w_2, ...w_j, ..., w_{k-1}, w_k\}$ is an ordered list of item sets corresponding to k visits where $t_{k-1} < t_k$.

Multi-hot Vector Representation

We start with a well-known data representation technique call multi-hot vector. This is based on the more common one-hot vector representation where only one element in an dictionary size vector can be 1 and rest are all zeros. This approach is useful when there is only one item at one time point. However, for the data we use, the EMR database includes one or more medical items per visit. Therefore, we set all items included in that visit to 1s and the rest are 0s. The input $x_t \in \mathbb{R}^{|\mathcal{L}|}$ is the multi-hot vector representation of v_t for the t-th time point. Figure 5.5 illustrates the conversion of item sets to multi-hot vectors where $|\mathcal{L}| = 7$ and three visits of one patient. Each vector (row of the matrics) is the summary of a patient visit input to the model.

Embedding Based Feature Vectors

Word embeddings are also a frequently used input representation where a lookup table is created for items and their representations. Then each item goes through this layer before entering the NN model. In this approach, a user can decide the output size of the embedding vector. The main goal of this approach is to decrease the sparsity by selecting a dimension smaller than the number of unique items. As the model is trained, the embeddings are updated according to the backpropagated errors that is computing based on the training objective. The training process typically allows

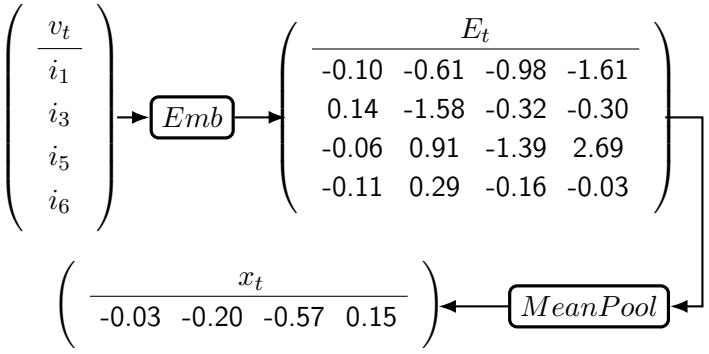


Figure 5.6: EMR embedding layer and mean pooling

the model to learn similarities of items and help the model improve its prediction accuracy. The embedding layer will output a fixed size vector for each item given as input. The embedding vectors of each individual item are combined using mean pooling approach because a visit is the combination of multiple items. Say, for a visit $v_t = (i_1, i_3, i_5, i_6)$ at time point t, e is the embedding size and the output is a matrix $E_t \in \mathbb{R}^{|v_t| \times e}$ as shown in Figure 5.6 (with $e = 4$). Then, mean pooling approach decreases the dimension and creates input for model as

$$x_t = \frac{E_t^1 + E_t^1 + \cdots + E_t^{|v_t|}}{|v_t|}. \tag{5.1}$$

This model is introduced in the DeepCare model to feed embedding vectors into prediction task with RNN based methods using the EMR database. While previous methods use diagnosis, procedure, and medication data all together as input, DeepCare separated diagnosis codes from others. However, is still combine medication and procedure codes and does not use any demographics information of the patient. After altering standard LSTM cell to handle these separate inputs, they showed performance gains in prediction. We only employ mean pooling approach since they

conclude that this approach gives better performance than other two approaches: max pooling and normalized sum pooling.

Concatenate Mean Pool Embedding

The problem with multi-hot vector representation is that size will grow with the unique item set size which makes it too sparse. Using embedding layer will solve the sparsity issue but still considers all medical items as semantically similar. An EMR dataset has multiple types of data and each type needs a different part in the input vector that we give to the model. Hence, we separate each medical data type into a new group. Then, each type goes through its own embedding and finally all type-specific embedding vectors are concatenated to generate the final input to the model. Let split medical items into 6 subgroups: diagnoses, procedures, medications, age at the visit, bmi at the visit, and tobacco usage information at that visit as $\mathcal{L}_{icd}$, $\mathcal{L}_{cpt}$, $\mathcal{L}_{med}$, $\mathcal{L}_{age}$, $\mathcal{L}_{bmi}$, and $\mathcal{L}_{tobac}$, respectively. Diagnoses, procedures, and medications data of a patient are vectors and we use mean pooling embedding approach separately. Each demographic information component of a patient is a scalar; therefore, the embedding layer will be enough without the mean pooling part.

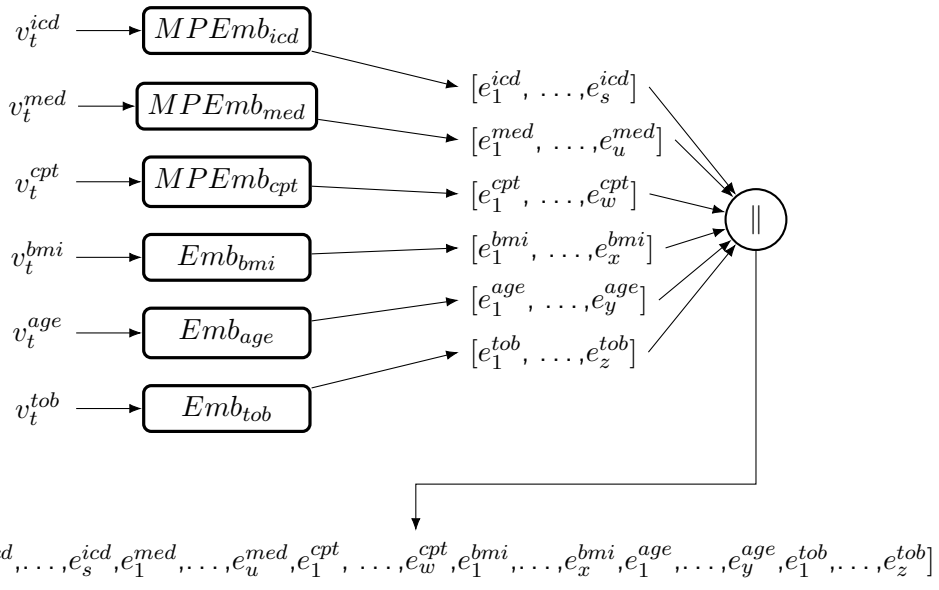


Figure 5.7: Concatenate embedding layer with mean pooling for input representation

Figure 5.7 illustrates the process of generating input vector from separate embedding vectors where $v_t^{icd} \in \mathcal{L}_{icd}$, $v_t^{med} \in \mathcal{L}_{med}$, $v_t^{cpt} \in \mathcal{L}_{cpt}$, $v_t^{bmi} \in \mathcal{L}_{bmi}$, $v_t^{age} \in \mathcal{L}_{age}$, and $v_t^{tob} \in \mathcal{L}_{tob}$ and s, u, w, x, y, and z are the embedding sizes respectively. In this

Figure, Embs are embedding nodes, MPEmbs are mean pooling nodes while $\|$ is the concatenation function for the vectors.

5.3.4 RETAIN Model

The RETAIN architecture (Choi et al., 2016b) mimics doctors' typical mode of operation. Doctors review a patient's medical history; by starting from the most recent one, they check patient records to understand what is going on with their health. RETAIN processes the longitudinal EMR data the same way. It uses two attention mechanisms where first one takes the input with the same visit order while the second attention mechanism reverses patient's historical data and start from the latest one and goes through the earliest patient visit. That is, the latest patient visit garners more attention of the model. The attention components of the model are based on GRU model and the general schema is shown in Figure 5.8. Each patient's medical record in the EMR database is converted to a multihot vector and this vector is given as input to the main structure. First, these multi-hot vectors go through the embedding layer which consists of a linear unit. Second, two attention mechanisms based on GRU use output of the embedding layer as input. One of them uses date in default order while the other one reverses the order before starting. Then, output of these three parts are combined and the final output is derived from them. The

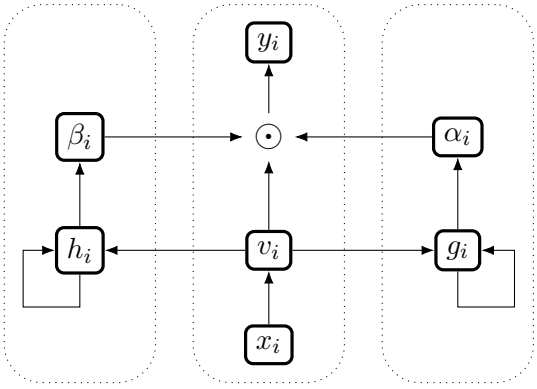


Figure 5.8: The RETAIN longitudinal EMR modeling architecture

mathematical equations for RETAIN are listed as:

$$v_i = W_{emb} x_i, \tag{5.2}$$

$$g_i, g_{i-1}, ..., g_1 = RNN_\alpha(v_i, v_{i-1}, \ldots, v_1), \tag{5.3}$$

$$e_j = w_\alpha^\top g_j + b_\alpha \quad for\ j = 1, \ldots, i, \tag{5.4}$$

$$\alpha_1, \alpha_2, \ldots, \alpha_i = Softmax(e_1, e_2, \ldots, e_i), \tag{5.5}$$

$$h_i, h_{i-1}, ..., h_1 = RNN_\beta(v_i, v_{i-1}, \ldots, v_1), \tag{5.6}$$

$$\beta_j = tanh(W_\beta h_j + b_\beta) \quad for\ j = 1, \ldots, i \tag{5.7}$$

$$c_i = \sum_{j=1}^{i} \alpha_j \beta_j \odot v_j, \tag{5.8}$$

$$\hat{y} = Softmax(W c_i + b). \tag{5.9}$$

Equation 5.2 is the first step where linear model is used to generate the embedding from multi-hot input vector and W_{emb} is the weight matrix of embedding layer. α values are generated by Equations 5.3, 5.4, and 5.5 while β values generated through Equations 5.6, and 5.7 using the output of the previous step. Then, a context vector is obtained by Equation 5.8 and prediction is made by Equation 5.9 using the context vector c_i. Cross-entropy loss function shown in Equation 5.10 where N is batch size is used as the training objective. This method is also proposed to predict future diagnoses of patients. However, we use a special case of the model where it predicts only one specific disease.

$$\mathcal{L}(x_1, \ldots, x_N) = -\frac{1}{N} \sum_{i=1}^{N} (y_i \log(\hat{y_i}) + (1 - y_i) \log(1 - \hat{y_i})), \tag{5.10}$$

where y_i is the ground truth Boolean value and $haty_i$ is the corresponding prediction probability from the architecture.

5.3.5 DeepCare Model

The DeepCare (Pham et al., 2017) model is created by modifying the LSTM cell in a way that separately accounts for each type of input codes to capture the therapeutic effects of interventions (medications and procedures) on diseases. Input to DeepCare consists of three parts: diseases, procedures, and medications, and CVG. At one time point or for a patient visit in this case, there are one or more medical codes. The embedding vectors are created for each code separately and need to be combined. In

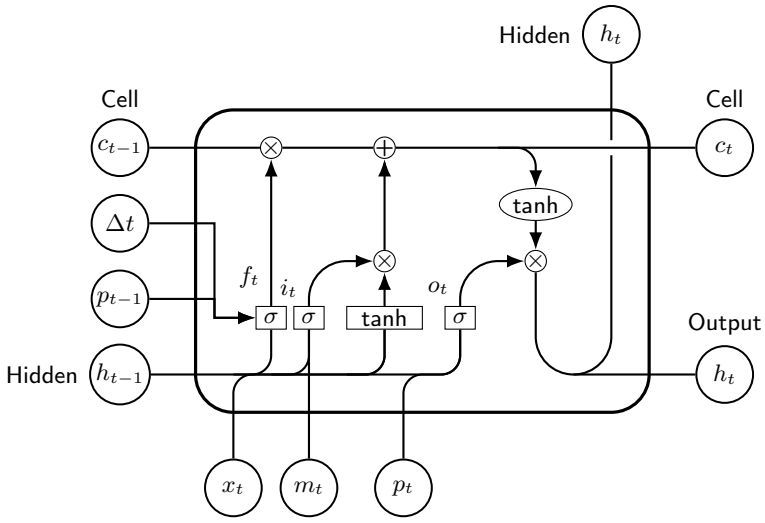


Figure 5.9: C-LSTM structure used in DeepCare model

this model, three different type of embeddings were tested and for simplicity we are using mean pooling approach, which generated the best prediction accuracy. Mean pooling is applied to diagnosis codes and the combination of procedure codes and medication codes separately. The mathematical equations for DeepCare model are listed as:

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + P_o p_t + b_o) \tag{5.11}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + Q_f q_{\Delta_{t-1:t}} + P_f p_{t-1} + b_f) \tag{5.12}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{5.13}$$

$$\tilde{c_t} = tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{5.14}$$

$$c_t = \sigma(f_t \odot c_{t-1} + i_t \odot \tilde{c_t}) \tag{5.15}$$

$$h_t = tanh(c_t) \odot o_t \tag{5.16}$$

In this Equations, $W_o, W_f, W_i, W_c, U_o, U_f, U_i, U_c, P_o, P_f$, and Q_f are weight matrices and b_o, b_f, b_i, b_c are the bias vectors that the model is training to learn. Equations 5.11, 5.12, and 5.13 are the output gate, forgot gate, and input gate respectively. Here, disease codes are the main inputs and affect all gates, procedure codes are used in output and forget gates. However, CVG data is only used by forget gate. An example third-degree forgetting CVG data representation is created as:

$$q_{\Delta_{t-1:t}} = \left(\frac{\Delta_{t-1:t}}{60}, \left(\frac{\Delta_{t-1:t}}{180} \right)^2, \left(\frac{\Delta_{t-1:t}}{365} \right)^3 \right) \tag{5.17}$$

where $\Delta_{t-1:t}$ is the CVG between visit t and $t-1$. In this chapter, we use a special case of the model with only one output. Binary cross-entropy used as the objective for training.

5.3.6 Two Level Hierarchical LSTM Model

In this section, we describe a new architecture that combines sequential contrast patterns with deep neural networks. The idea is to design a two-level hierarchical model, as shown in Figure 5.10. We feed a transaction to the first level. This level includes an embedding layer and LSTM cell running consecutively. Here, each SCP consists of multiple itemsets in temporal order, $S_k = c_1^k, c_2^k, \ldots, c_l^k$. Embedding layer using mean pool embedding reads these itemsets and outputs embedded vectors as shown in Figure 5.6. Embedding layer creates an embedded vector for each item. However, each itemset in our sequence consists of multiple items, so we average the embedded vector to derive the final embedding for the itemset. Mathematical functions for the first level LSTM model are listed as:

$$i_t^1 = \sigma(W_i^1 x_t + U_i^1 g_{t-1} + b_i^1), \tag{5.18}$$
$$f_t^1 = \sigma(W_f^1 x_t + U_f^1 g_{t-1} + b_f^1), \tag{5.19}$$
$$o_t^1 = \sigma(W_o^1 x_t + U_o^1 g_{t-1} + b_o^1), \tag{5.20}$$
$$\tilde{c_t}^1 = tanh(W_c^1 x_t + U_c^1 g_{t-1} + b_c^1), \tag{5.21}$$
$$c_t^1 = \sigma(f_t^1 \odot c_{t-1}^1 + i_t^1 \odot \tilde{c_t}^1), \tag{5.22}$$
$$g_t = tanh(c_t^1) \odot o_t^1. \tag{5.23}$$

Mathematical functions for the second level LSTM model are listed as:

$$i_t^2 = \sigma(W_i^2 g_t + U_i^2 h_{t-1}^{seq} + b_i^2), \tag{5.24}$$
$$f_t^2 = \sigma(W_f^2 g_t + U_f^2 h_{t-1}^{seq} + b_f^2), \tag{5.25}$$
$$o_t^2 = \sigma(W_o^2 g_t + U_o^2 h_{t-1}^{seq} + b_o^2), \tag{5.26}$$
$$\tilde{c_t}^2 = tanh(W_c^2 g_t + U_c^2 h_{t-1}^{seq} + b_c^2), \tag{5.27}$$
$$c_t^2 = \sigma(f_t^2 \odot c_{t-1}^2 + i_t^2 \odot \tilde{c_t}^2), \tag{5.28}$$
$$h_t^{seq} = tanh(c_t^2) \odot o_t^2. \tag{5.29}$$

At a high level, the first level LSTM cell is fed by embedded vectors of all code sets in an SCP. Its output is its summary, g_t, that goes to the second layer. The second layer outputs h_t as final output of this model. Here, embedding dimension is m and

hidden layer size is p. Each U parameter is from $\mathbb{R}^{m \times p}$ and each W parameter is from $\mathbb{R}^{m \times p}$ and the bias parameters b are from $\mathbb{R}^m$. c_t^j is the cell state of the corresponding layer while g_t is the hidden layer of the first level and h_t^{seq} is the hidden layer of the second level LSTM.

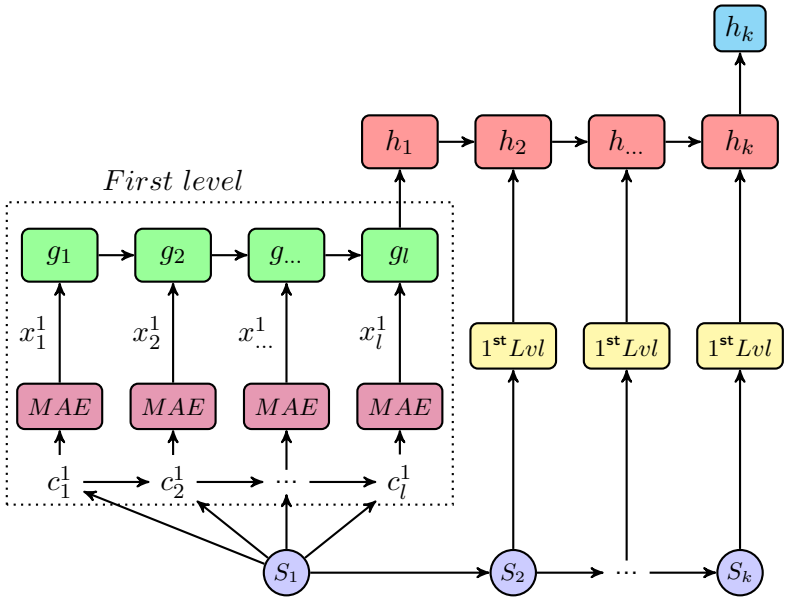


Figure 5.10: 2-level hierarchical SCP based neural architecture for predictive modeling through longitudinal EMRs

5.3.7 Combining Our Model with V-LSTM

In this section we describe a simple combination architecture that derives from the V-LSTM model (Section 2.7.1) and the 2lvl-H-LSTM model (Section 5.3.6). For each patient, we have both SP database and the input vectors. Intuitively we can run both models simultaneously and merge the intermediate outputs before the prediction is made in the final layer. To that end, the output vectors of both models are added and the sum is fed to the dropout layer (see Figure 5.11). Afterwards, the output of dropout layer is given as input to the ReLU unit. Finally, we added 2-level linear layer on top of these layers to decrease the size to a single scalar since we are interested in binary classification. The final prediction is made using a sigmoid unit.

Since we unbalanced EMR, we calculated the loss for the training by weighted cross-entropy loss function shown as:

$$\mathcal{L}(x_1, \ldots, x_N) = -\frac{1}{N} \sum_{i=1}^{N} (w_p y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)) \tag{5.30}$$

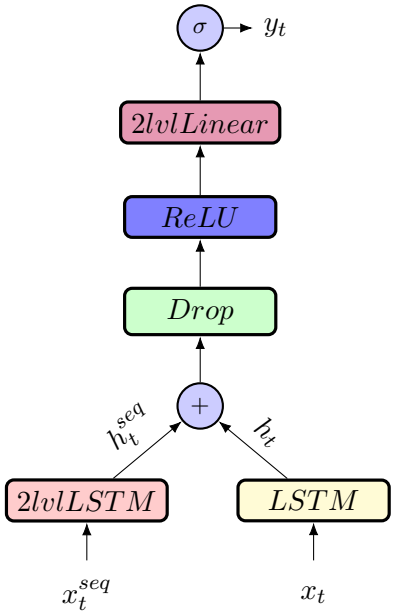


Figure 5.11: Embedding layer and mean pooling

where N is the batch size and $w_p = \frac{|EMR_n|}{|EMR_p|}$, the ratio of negative samples size over positive sample size.

5.4 Results and Discussion

We first discuss the performance of our hierarchical model in comparison with the results of other baseline models introduced earlier: Doctor AI, RETAIN, V-LSTM, and DeepCare. For each of these models, we also generated a concatenate embedding version by changing the embedding layer.

5.4.1 Experimental Setup

In this study, we designed a model which is the combination of two models where first one is an 2lvl-H-LSTM taking SCP based database and second one is the V-LSTM taking concatenated mean pool embedding () vectors of each input type. We used the EMR dataset collected from 14,294,463 visits from 1,120,330 patients from the UKY medical center and affiliated clinics. Before starting our experiments, we created a cohort using 6 months of washout period, 12 months of max-CVG, 12 months of prediction window, and patients with at least 30 visits to the medical center. These database statistics are given in Table 5.5. We mine the top 10,000 SCPs according to the relative risk score for our combined model using frequency threshold $\sigma = 100$ and maximum gap allowed between consecutive itemsets in a sequence $gap = 2$. After

applying these restrictions, we obtained SCPs with relative risk value between 2.94 and 0.66. Relative risk higher than 1 indicates that the risk of a patient to develop DD is increased by a given pattern while values below 1 indicate that the sequence decreases such risk. Finally, the value of 1 means no relationship. We also run the same experiment with 5% ($\sigma = 176$) frequency. At the beginning, our data contains 679 unique medical codes. After running SCPM application, 197 and 268 of them are remained for support of 100 and 176 inputs respectively; Only 71 and 101 of them are entered as input to the LSTM unit after converting our database to a SP based one, respectively.

Measure	Value
Total patient count	1,120,330
Patients diagnosed with depression	73,080
Total visit count	14,294,463
Average number of Visits	12.75
Washout period	6 Months

Table 5.5: The EMR database statistics used for predictive modeling for depressive disorders

5.4.2 Models

Here we compare the output of ten models to understand the performance of the prediction task. We designed two new models for this study: one is the 2lvl-H-LSTM and the other one is the combined model. We have output of our hierarchical model with the support of 176 and 100. Four of them are previously designed models by other researchers and the remaining four are the models we modified changing embedding layer to a concatenate embedding approach. Then, we run each test 10 times with 20 epochs (*i.e.*, 200 total runs over the whole training data) and changed batch size depending on the model (to stick with the original implementations). The final scores presented in this section are the averages of ten runs of each model. All models used in this study are listed below.

- The first baseline is the Doctor AI with two hidden layers. Since we use grouped medical codes, we set hidden layer size to 200, learning rate to 0.0001 and the batch size to 32. This model uses a multi-hot vector as embedding.
- The second baseline of our study is RETAIN which uses hidden layer size of 128 for both GRU units and the batch size of 100. This model also uses a multi-hot vector as embedding.

- Next baseline is DeepCare model with hidden layer size of 100 and the batch size of 64 patients. In this method, we used diagnosis codes separated from the procedure and medication codes as explained in the original work. The embedding size of this model is 200 for each of diagnosis codes and the combination of procedures and medications. Here, we employed MPEmb technique (from Section 5.3.3).

- We also designed a V-LSTM model as a baseline with batch size of 64 and hidden layer size of 100, parameters we chose after trying a few settings. This baseline uses only diagnosis, medication and procedure codes as input.

- First part of our original model is a 2lvl-H-LSTM model. We computed the result of this model to compare with the baseline models. In this part, we used hidden layer size of 50 for both layers with the batch size of 32.

- Next model is the combination of our 2lvl-H-LSTM and the V-LSTM models. In the embedding layer, we separated each type of data and used different embedding sizes for each one according to the number of unique inputs. We employed six different input types which are diagnoses, medications, procedures, tobacco status, patient age at the visit, and patient BMI at the visit with embedding sizes of 50, 50, 50, 5, 5, 5 respectively. We run MPEmb layer for each input type and concatenate these embedding vectors to create the final input.

- Finally there are four modified baseline models having the same hidden and batch sizes as the original versions with concatenated MPEmb vectors (from Section 5.3.3).

5.4.3 Comparing the Results of Models

To measure the performance differences due to input representation, we run each base model and the concatenated embedding version for 32 different cohorts (from Table 5.3). Table 5.6 shows all the mean differences between F1-scores, precision scores, and recall scores of an original baseline model and the corresponding model with the concatenated MPEmb input version of it. The difference is taken as the concatenated MPEmb version's score minus original version. Therefore, in this table, positive values mean the modified version has a better score. Our results show that V-LSTM$_{ConCat}$ model outperforms the original model in 30 of 32 cases (from Table 5.7) with the average of 6.1% improvement while Doctor AI$_{ConCat}$ model has 4.3% average

Table 5.6: Comparison of the results of models with different input representations: the concatenated MPEmb version versus original version. Mean is taken over the differences from model performances on all 32 cohorts.

V-LSTM			
Measure	F1-Score Diff.	Precision Diff.	Recall Diff.
Mean	**0.06187**	**0.0733**	0.06187
Standard Error	0.00429	0.00519	0.00429
Median	0.06255	0.0668	0.06255
Minimum	0.0043	0.02432	0.0043
Maximum	0.09657	0.13263	0.09657
Sum	1.97982	2.34567	1.97982
Confidence Level (95.0%)	0.00876	0.01059	0.00876
Doctor AI			
Mean	0.04329	0.03181	**0.08848**
Standard Error	0.00481	0.00586	0.02237
Median	0.03967	0.03770	0.10463
Minimum	-0.01945	-0.03395	-0.27395
Maximum	0.10048	0.10718	0.30180
Confidence Level (95.0%)	0.00982	0.01196	0.04563
RETAIN			
Mean	0.00039	0.00827	-0.10802
Standard Error	0.00318	0.00421	0.02500
Median	0.00027	0.00107	-0.06584
Minimum	-0.04328	-0.02236	-0.50192
Maximum	0.04964	0.08975	0.03180
Confidence Level (95.0%)	0.00648	0.00859	0.05101
DeepCare			
Mean	0.05369	0.06149	0.02541
Standard Error	0.00411	0.00461	0.00860
Median	0.05657	0.06524	0.02603
Minimum	0.00846	0.00714	-0.08679
Maximum	0.09572	0.10366	0.13307
Confidence Level (95.0%)	0.00837	0.00940	0.01754

improvement and outperforms in 31 scenarios (from Table 5.8). DeepCare$_{ConCat}$ is the model which yields better results (from Table 5.10) in all cases with the biggest mean improvement of 5.3% on F1-score. On the other hand, the concatenated version of RETAIN only outperforms in 17 of 32 cohorts (from Table 5.9) and the mean average F1-score difference is 0.00039, which means that there is very slight improvement on average F1-score of the original model and the modified version.

From Tables 5.7–5.10, our results also indicate that max-CVG of 12 months improves the F1-score while applying 3 and 6 months restrictions decreases the performance. That is, these restrictions may be too strong given they result in a significant decrease in the data size. Another observation from the results is that the prediction performance increases when the minimum number of patient visits restriction is increased. We noticed that the RETAIN model gives some unexpected results.

It outputs the worst results in all cases despite taking the longest running time of all model. That outcome makes the RETAIN model unreliable among others for our data. The validation part of our experiment chooses very low thresholds for the classification part which results to the model to choose most patients as positive. Because of this behaviour, the recall score is always high while the precision score and F1-score are very low. Another observation is that the prediction horizon window of 6 months gives better performance in most cases as expected because choosing 12 months decreases observable history of each patient. This behaviour aligns with the minimum number of visits restriction.

Table 5.7: Comparison of results for original V-LSTM model without demographics with the V-LSTM model using ConCat Embedding across all 32 cohorts

				V-LSTM			V-LSTM ConCat Embedding		
Min Visit	CVG	Obs Win	Washout	F-Score	Presicion	Recall	F-Score	Presicion	Recall
0	0	6	6	0.26702	0.20403	0.3913	0.36359	0.28685	0.49814
10	0	6	6	0.35642	0.28336	0.48366	0.44417	0.36178	0.57656
20	0	6	6	0.46181	0.39685	0.55577	0.53105	0.48865	0.58555
30	0	6	6	0.52333	0.48742	0.57049	0.57942	0.55458	0.60932
0	3	6	6	0.27053	0.16042	0.89082	0.31564	0.21232	0.64388
10	3	6	6	0.27072	0.15821	0.9449	0.31106	0.22456	0.5551
20	3	6	6	0.31431	0.19289	0.87604	0.35906	0.27364	0.59271
30	3	6	6	0.36936	0.24627	0.76813	0.40007	0.29025	0.66703
0	6	6	6	0.31255	0.20672	0.65	0.37698	0.27039	0.63562
10	6	6	6	0.34245	0.23473	0.653	0.40527	0.29449	0.68134
20	6	6	6	0.39753	0.29098	0.64293	0.45852	0.38942	0.57049
30	6	6	6	0.496	0.46794	0.55606	0.54152	0.52504	0.57008
0	12	6	6	0.34033	0.25132	0.53114	0.42402	0.3706	0.49961
10	12	6	6	0.36309	0.26521	0.58071	0.45728	0.39784	0.54294
20	12	6	6	0.45222	0.37801	0.56654	0.51227	0.49135	0.54086
30	12	6	6	0.52072	0.47259	0.58998	0.57284	0.53729	0.61808
0	0	12	6	0.26376	0.20222	0.38602	0.34886	0.26341	0.52066
10	0	12	6	0.33067	0.24252	0.52385	0.42135	0.33503	0.56885
20	0	12	6	0.43429	0.36353	0.54383	0.50579	0.44744	0.58398
30	0	12	6	0.50669	0.4883	0.5298	0.57049	0.53552	0.61333
0	3	12	6	0.3003	0.17815	0.96038	0.3046	0.20614	0.6566
10	3	12	6	0.29867	0.17593	0.98868	0.33843	0.23205	0.69057
20	3	12	6	0.31497	0.19681	0.83396	0.32731	0.22124	0.68491
30	3	12	6	0.33751	0.208	0.89811	0.36288	0.23232	0.85849
0	6	12	6	0.29585	0.19343	0.6589	0.37685	0.25986	0.72192
10	6	12	6	0.33762	0.23864	0.66186	0.38019	0.27094	0.68454
20	6	12	6	0.33836	0.26214	0.57923	0.42014	0.31177	0.67324
30	6	12	6	0.39871	0.28679	0.67333	0.45819	0.35625	0.66037
0	12	12	6	0.31961	0.23803	0.50031	0.40079	0.33315	0.51303
10	12	12	6	0.33715	0.23716	0.5879	0.42587	0.35801	0.53153
20	12	12	6	0.38979	0.28789	0.63266	0.48538	0.40089	0.61989
30	12	12	6	0.48567	0.41827	0.58461	0.54795	0.52736	0.57241

Finally, we present the effect of combining SCPs with a deep learning model. We

Table 5.8: Comparison of results for original Doctor AI model with Doctor AI model using ConCat embedding across all 32 cohorts

				Doctor AI Original			Doctor AI ConCat Embedding		
Min Visit	CVG	Obs Win	Washout	F-Score	Presicion	Recall	F-Score	Presicion	Recall
0	0	6	6	0.26548	0.21814	0.38692	0.36597	0.27061	0.56893
10	0	6	6	0.36315	0.29494	0.48727	0.44594	0.34560	0.63562
20	0	6	6	0.46895	0.39470	0.62270	0.52532	0.44762	0.65448
30	0	6	6	0.54184	0.48979	0.65083	0.57596	0.52294	0.67754
0	3	6	6	0.27374	0.18876	0.60918	0.35006	0.25195	0.74591
10	3	6	6	0.27393	0.19584	0.73061	0.34015	0.23194	0.75816
20	3	6	6	0.29826	0.23461	0.81562	0.33738	0.31368	0.54166
30	3	6	6	0.38813	0.24539	0.94175	0.41171	0.28899	0.89560
0	6	6	6	0.36317	0.27491	0.60251	0.39128	0.29879	0.73698
10	6	6	6	0.38166	0.28904	0.58640	0.39419	0.26158	0.85276
20	6	6	6	0.43358	0.35242	0.61707	0.47380	0.40584	0.68512
30	6	6	6	0.44469	0.37888	0.55390	0.51199	0.44491	0.76765
0	12	6	6	0.36884	0.32867	0.44400	0.43581	0.34348	0.64423
10	12	6	6	0.38231	0.29470	0.56642	0.42058	0.36542	0.66371
20	12	6	6	0.49470	0.42204	0.60813	0.51009	0.44234	0.70658
30	12	6	6	0.53180	0.44020	0.71332	0.58234	0.54739	0.69204
0	0	12	6	0.26272	0.19681	0.42113	0.34949	0.24294	0.63382
10	0	12	6	0.33529	0.26635	0.49109	0.41987	0.32736	0.60192
20	0	12	6	0.45671	0.40025	0.54285	0.49950	0.39587	0.69004
30	0	12	6	0.52188	0.46151	0.62489	0.54466	0.46161	0.72032
0	3	12	6	0.30143	0.18803	0.83018	0.32177	0.21931	0.78537
10	3	12	6	0.30831	0.19231	0.84716	0.28885	0.19061	0.71886
20	3	12	6	0.31031	0.18425	0.98742	0.34357	0.22356	0.89622
30	3	12	6	0.33674	0.20246	1	0.38998	0.25596	0.92075
0	6	12	6	0.36277	0.26350	0.63801	0.41243	0.31310	0.72945
10	6	12	6	0.35439	0.29180	0.53058	0.41326	0.31036	0.75326
20	6	12	6	0.41056	0.32628	0.64295	0.44472	0.37467	0.66302
30	6	12	6	0.44313	0.33611	0.67444	0.44575	0.30216	0.89703
0	12	12	6	0.35557	0.27910	0.54358	0.39370	0.27296	0.74629
10	12	12	6	0.35414	0.31613	0.45552	0.40788	0.29203	0.75732
20	12	12	6	0.43501	0.32181	0.69615	0.44390	0.32440	0.82892
30	12	12	6	0.50546	0.41805	0.67229	0.52202	0.41579	0.79692

show that creating a hybrid model has a positive effect to the prediction task. As shown in Table 5.11*, our model outperformed the state-of-the-art DeepCare model with 6% higher F1-score for the original version and 4% for the concatenated embedding versions and RETAIN model always gives the lowest score. Although SCPM removes 60%-70% of total input, using 2lvl-H-LSTM solely has comparable predictive power to the other models and even better results than the RETAIN model. When we compare 2lvl-H-LSTM model with two different frequencies, we identified that the lower frequency experiment gives better result (from an F-score perspective) but

*It is important to note here that each of these scores is an average of performances from ten different models (of the same architecture). As the average is taken for F-score, precision, and recall separately, the displayed F-score may not compute using the shown recall and precision means.

Table 5.9: Comparison of results for original RETAIN model with RETAIN model using ConCat embedding across all 32 cohorts

				Retain Original			RETAIN ConCat Embedding		
Min Visit	CVG	Obs Win	Washout	F-Score	Presicion	Recall	F-Score	Presicion	Recall
0	0	6	6	0.13763	0.07520	0.81033	0.10242	0.05790	0.44437
10	0	6	6	0.19786	0.11159	0.87228	0.19816	0.11231	0.84133
20	0	6	6	0.26851	0.15713	0.92236	0.30204	0.23168	0.44426
30	0	6	6	0.33167	0.20358	0.89439	0.38131	0.25411	0.78768
0	3	6	6	0.26866	0.15594	0.96939	0.26080	0.15251	0.90000
10	3	6	6	0.27412	0.15901	0.99286	0.25803	0.15218	0.85000
20	3	6	6	0.31003	0.18471	0.96458	0.29837	0.17883	0.90000
30	3	6	6	0.36016	0.22170	0.95934	0.35619	0.22331	0.88242
0	6	6	6	0.20552	0.11659	0.86644	0.20678	0.11733	0.87009
10	6	6	6	0.21958	0.12574	0.86567	0.21773	0.12484	0.85092
20	6	6	6	0.31268	0.18883	0.90976	0.31446	0.19006	0.91098
30	6	6	6	0.38249	0.24098	0.92668	0.38719	0.24994	0.85957
0	12	6	6	0.17198	0.09600	0.82463	0.17467	0.09751	0.83702
10	12	6	6	0.21550	0.12297	0.87085	0.21709	0.12389	0.87677
20	12	6	6	0.29521	0.17811	0.86179	0.30041	0.18124	0.87879
30	12	6	6	0.36950	0.23152	0.91467	0.38749	0.25719	0.78833
0	0	12	6	0.14001	0.07658	0.81543	0.09673	0.05422	0.44885
10	0	12	6	0.19590	0.11017	0.88318	0.19753	0.11155	0.86195
20	0	12	6	0.25481	0.15124	0.80843	0.25164	0.15959	0.67368
30	0	12	6	0.31198	0.18876	0.89858	0.32399	0.27852	0.39665
0	3	12	6	0.28587	0.16782	0.96415	0.27719	0.16508	0.86415
10	3	12	6	0.29559	0.17343	1.00000	0.28691	0.17260	0.85094
20	3	12	6	0.30788	0.18299	0.96981	0.30972	0.18679	0.90755
30	3	12	6	0.33224	0.20030	0.97358	0.33187	0.20266	0.91698
0	6	12	6	0.23703	0.13645	0.90171	0.23770	0.13717	0.89041
10	6	12	6	0.23988	0.13838	0.90034	0.24013	0.13875	0.89210
20	6	12	6	0.29312	0.17360	0.94120	0.29109	0.17407	0.88873
30	6	12	6	0.34554	0.21336	0.90852	0.33295	0.20777	0.83889
0	12	12	6	0.19149	0.10790	0.84984	0.19720	0.11102	0.88165
10	12	12	6	0.21514	0.12254	0.88068	0.21152	0.12087	0.85042
20	12	12	6	0.27784	0.16405	0.90746	0.26595	0.17454	0.59684
30	12	12	6	0.34298	0.20972	0.94138	0.38565	0.25152	0.83128

the higher frequency one is better from a recall point of view. Here, decreasing the minimum frequency also helps to capture more predictive sequences. This behaviour is expected since higher frequency yields lower number of items, number of patterns to compare, and quality patterns. Also, our model gives a better trade-off between precision and recall with a four point difference. However, other baseline models tend to heavily prioritize one over the other. The smallest difference between precision and recall in our baseline experiments is the V-LSTM model with a nearly 10 point gap.

Table 5.10: Comparison of results for original DeepCare model with DeepCare model using ConCat embedding across all 32 cohorts

				DeepCare Original			DeepCare ConCat Embedding		
Min Visit	CVG	Obs Win	Washout	F-Score	Presicion	Recall	F-Score	Presicion	Recall
0	0	6	6	0.26998	0.20952	0.38340	0.36570	0.29268	0.48907
10	0	6	6	0.35959	0.28895	0.47918	0.45467	0.39261	0.54138
20	0	6	6	0.48056	0.42000	0.56268	0.54561	0.50091	0.60053
30	0	6	6	0.55477	0.53700	0.57544	0.59959	0.57805	0.62390
0	3	6	6	0.29889	0.21791	0.49184	0.35804	0.25504	0.61531
10	3	6	6	0.32350	0.22606	0.59694	0.35869	0.25748	0.61224
20	3	6	6	0.34217	0.24926	0.56979	0.38016	0.28185	0.59896
30	3	6	6	0.44995	0.35014	0.64615	0.46366	0.35729	0.68352
0	6	6	6	0.35917	0.26957	0.54292	0.41606	0.34611	0.52740
10	6	6	6	0.38037	0.27927	0.60415	0.42977	0.33605	0.60230
20	6	6	6	0.45216	0.36528	0.59732	0.48616	0.42357	0.57341
30	6	6	6	0.52346	0.47688	0.58733	0.54922	0.50341	0.61078
0	12	6	6	0.36118	0.29774	0.46180	0.44046	0.38510	0.51843
10	12	6	6	0.38974	0.32072	0.50115	0.46353	0.40929	0.53645
20	12	6	6	0.47978	0.42345	0.55731	0.53000	0.50057	0.56581
30	12	6	6	0.54369	0.51179	0.58120	0.58566	0.58068	0.59163
0	0	12	6	0.26389	0.20502	0.37485	0.35730	0.27614	0.50793
10	0	12	6	0.33909	0.25787	0.49952	0.43063	0.35524	0.55035
20	0	12	6	0.45483	0.39157	0.54390	0.51451	0.46833	0.57253
30	0	12	6	0.52650	0.48746	0.57469	0.58287	0.54845	0.62234
0	3	12	6	0.33591	0.22363	0.68679	0.36348	0.26447	0.60000
10	3	12	6	0.33581	0.22536	0.67736	0.36105	0.26078	0.62453
20	3	12	6	0.35737	0.25335	0.65472	0.36583	0.27200	0.59434
30	3	12	6	0.38075	0.29139	0.56604	0.41743	0.30947	0.65660
0	6	12	6	0.36457	0.28003	0.53493	0.42136	0.33767	0.56849
10	6	12	6	0.36987	0.29285	0.51168	0.43116	0.37081	0.52509
20	6	12	6	0.41083	0.32832	0.55704	0.46783	0.39849	0.57359
30	6	12	6	0.45858	0.36258	0.62778	0.49707	0.41362	0.62852
0	12	12	6	0.33639	0.28055	0.42388	0.41688	0.36592	0.48613
10	12	12	6	0.37428	0.30616	0.48567	0.44093	0.39968	0.49682
20	12	12	6	0.43986	0.35422	0.58994	0.50282	0.44855	0.57480
30	12	12	6	0.51694	0.43223	0.64815	0.55462	0.49383	0.63547

5.5 Conclusion

In this chapter, we employed the EMR database containing 14.3M patient visits for 1.12M patients. This study involved two main parts: in the first part, we created multiple test cases and applied baseline models as well as change input embedding layers of these models to compare the improvements; in the second part we designed our novel SCP based model and applied one of these 32 use cases to see the improvements in prediction. For the first task, we created 32 different scenarios and outlined four baseline models previously studied in this area for the future diagnosis prediction task. We applied these models to each use case we created as well as the concatenated embedding versions. Our results show that changing the input improves

Table 5.11: Comparing results using washout period: 6 months, max-CVG: 12 months, prediction horizon window: 12 months, and min patient visit size: 30, with our hierarchical and the combined models.

Model	F1-score	Precision	Recall
V-LSTM$_{N}oDem$	0.46169	0.31963	0.87931
Doctor AI$_{org}$	0.50546	0.41805	0.67229
RETAIN$_{org}$	0.34298	0.20972	0.94138
DeepCare$_{org}$	0.51694	0.43223	0.64815
V-LSTM$_{ConCat}$	0.52731	0.53723	0.63571
Doctor AI$_{ConCat}$	0.52202	0.41579	0.79692
RETAIN$_{ConCat}$	0.38565	0.25152	0.83128
DeepCare$_{ConCat}$	0.55462	0.49383	0.63546
2lvl-H-LSTM$_{\sigma=176}$	0.40730	0.27240	0.82998
2lvl-H-LSTM$_{\sigma=100}$	0.42772	0.32989	0.63610
Combined model	**0.59087**	**0.57238**	0.61453

performance in most cases on every model. For the second part, we chose a cohort with max-CVG of 12 months, washout period of 6 months, prediction window of 12 months, and minimum number of patient visits to the facilities as 30. In order to predict a future condition of a patient, we designed a combined two component model where one part uses the V-LSTM and other part uses a two level hierarchical LSTM model with sequential database as input. We added the resulting vectors of these two models before making the prediction in a hybrid neural architecture. We showed that our model outperformed the best baseline model by 7.5% in F1-score when original embedding is employed and by 3.5% in F1-score when concatenated embedding is employed.

Chapter 6 Conclusion

Rapid adoption of EMRs incentivized by the federal government has created new affordances in secondary analyses and applications of EMR data beyond individual patient care and operational functionality. In this dissertation, we studied three possible applications using the EMR database from the UKHealthCare system with mental disorders as use-cases for each of them. Despite this mental health focus, our methods are readily applicable to any target chronic condition of interest. More specifically, these are our contributions.

- **Interestingness measures for association rule mining (ARM).** In Chapter 3, we ranked ARs using more than 40 interestingness measures including a few measures that we created for this study. We also obtained manually assigned novelty scores from our domain expert. Then, we ranked ARs according to the statistical strength and novelty. Our experiments surface groups of interestingness measures that weight rule novelty and statistical strength in contrasting ways, offering new insights for end users in identifying interesting rules.

- **Toward causal association rule (CAR) mining.** In Chapter 4, we glean CARs from the EMR database using matched fair datasets to calculate statistical strength after accounting for confounders. Then, we gather causality (biomedical plausibility) scores manually assigned by two domain experts. Comparing the scores showed that statistical strength on fair datasets using ML techniques align with domain expert scores for anxiety disorders and depressive disorders use cases.

- **Predictive modeling with contrast patterns and neural networks.** In Chapter 5, we designed a DL model to predict the first diagnosis of a medical condition using longitudinal EMRs. This design includes two parts where the first part is a traditional LSTM model and the second part enhances that model by sequential contrast patterns. Results shows that our model outperforms baseline models with at least 4 % improvement in F1-score. We also determined the benefit of code-type specific input representations as part of this study.

Limitations and Future Work

There are a few limitations in using an EMR database in ML applications. The massive scale (in terms of (1). number of unique variables and (2). numbers of patients and patient visits) makes most methods intractable for EMRs. In our experience, the current state-of-the-art in conventional ARM approaches including those that use the MapReduce paradigm do not scale well to very large datasets. The EMR databases contain clinical text beside the medical codes, which we have not used in our methods in this dissertation. We identify the following future opportunities to extend the work proposed in this dissertaiton.

- Developing numerical similarity metrics between patients is a recent trend with applications in decision making, predictive modeling (nearest neighbor methods), cohort selection, and phenotyping (Li et al., 2015b). Once patients are represented in $\mathbb{R}^d$, patient vectors will be compared using metrics such as cosine similarity or more task specific metrics learned in a supervised manner. Using patient LEMR vectors, we can use a basic cosine similarity model that takes as input a patient ID and returns a ranked list of most similar patients. An important goal of designing better patient representations and using them in similarity computation is to enable specialty clinicians to identify most similar patients to an incoming new patient in making choices about treatment options.
- With rapid rise in patients with multiple chronic comorbidities, we aim to gain specific insights via discriminative sequential patterns that occur temporally between the first diagnoses of two chronic conditions. An associated task is to identify subtypes within the group of patients that have the chronic condition pair. Here, we can impose a washout period for the first condition in the pair and a minimum gap of between both conditions. Contrast pattern mining methods can be used to identify discriminative sequential patterns that occur in the gap period between the diagnoses when compared with patients who just have one of the conditions in the pair.
- Finally, we plan to move toward associative classification (Yin and Han, 2003) approaches for specific chronic diseases and for designing new classification features for extracting coded information (Kavuluru and Lu, 2014; Kavuluru et al., 2015) as a further step from ARM.

Abbreviations

2lvl-H-LSTM 2-level hierarchical LSTM. 68–71, 74, 77

AHRQ Agency for Healthcare Research and Quality. 23

AIRF average inverse rule frequency. 24, 25

AR association rule. 19–22, 24, 31, 78

ARM association rule mining. 19–23, 31, 32, 79

BMI body mass index. 33, 37, 38, 50, 56

BN bayesian network. 33–35

C-LSTM Care Long Short Term Memory. 55

CA causal association. 5, 33–35, 39, 48

CAR causal association rule. 5, 34–36, 38, 39, 43–46, 48–50, 78

CDC Centers for Disease Control and Prevention. 22

ConSGapMiner Contrast Sequences with Gap Miner. 58

COPD chronic obstructive pulmonary disease. 53

CPM contrast pattern mining. 14

CPT Current Procedural Terminology. 56

CPU central processing unit. 54, 55

CSS clinical classifications software. 37, 56

CVG consecutive visit gap. 57, 65–67, 69, 72–77

DD depressive disorders. 57, 70

DL deep learning. 16, 78

EMR electronic medical record. 14, 19, 20, 31, 34–39, 42–45, 48, 50, 54–56, 58, 61–64, 68, 69, 76, 78, 79

FD fair dataset. 5

FIM frequent itemset mining. 54

GPU graphical processing unit. 54, 55

GRU gated recurrent unit. 17, 18, 55, 64, 70

HCUP Healthcare Cost and Utilization Project. 23, 24, 29–31, 37

ICD-10 International Classification of Diseases 10^{th} revision. 56

ICD-9 International Classification of Diseases 9^{th} revision. 23, 24, 53–56

ICD-9-CM International Classification of Diseases, Clinical Modification, 9^{th} revision. vi, 8, 9, 22, 37, 38

KNN k-nearest neighbour. 55

LCM Linear-time Closed item set Miner. 23

LR logistic regression. 55

LSTM long short term memory. 17, 55, 58, 62, 65, 67, 68, 70, 77, 78

MeSH medical subject headings. 29

ML machine learning. 79

MLP multilayer perceptron. 55

MPEmb mean pool embedding. 71

NDCG normalized discounted cumulative gain. 27

NDCN normalized discounted cumulative novelty. 27–29

NDCO normalized discounted cumulative ORLB. 27–29

NN neural network. 62

NSCLC Non-Small Cell Lung Cancer. 21

OR odds ratio. 25, 54

ORLB odds ratio lower bond. 5, 25–32, 36, 41, 43–46, 49, 50

RCT randomized control trial. 33–35

ReLU rectified linear unit. 68

RETAIN REverse Time AttentIoN. 55, 64, 65, 69, 70, 72–74, 77

RNN recurrent neural network. 15–17, 52, 53, 55, 60, 62

SCP sequential contrast pattern. 14, 15, 53, 54, 58–60, 67, 69, 70, 73

SCPM sequential contrast pattern mining. 70, 74

SP sequential pattern. 14, 15, 54, 58, 60, 68, 70

SPM sequential pattern mining. 53–55, 58

UKY the University of Kentucky. 5, 21, 22, 36, 56, 69

V-LSTM the vanilla long short term memory. 16, 17, 53, 58, 68, 69, 71–73, 75, 77

Bibliography

[1] O. Abar, R. J. Charnigo, A. Rayapati, and R. Kavuluru. "On Interestingness Measures for Mining Statistically Significant and Novel Clinical Associations from EMRs". In: *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM. 2016, pp. 587–594.

[2] S. Abuse et al. "Results from the 2005 national survey on drug use and health: national findings". In: *http://www. oas. samhsa. gov/nsduh/2k5nsduh/2k5Results. pdf* (2006).

[3] R. Agrawal and R. Srikant. "Fast Algorithms for Mining Association Rules in Large Databases". In: *Proceedings of the 20th International Conference on Very Large Data Bases*. Morgan Kaufmann Publishers Inc. 1994, pp. 487–499.

[4] J. Ayres, J. Flannick, J. Gehrke, and T. Yiu. "Sequential pattern mining using a bitmap representation". In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2002, pp. 429–435.

[5] T. A. Blakely, S. C. Collings, and J. Atkinson. "Unemployment and suicide. Evidence for a causal association?" In: *Journal of Epidemiology & Community Health* 57.8 (2003), pp. 594–600.

[6] S. E. Brossette, A. P. Sprague, J. M. Hardin, K. B. Waites, W. T. Jones, and S. A. Moser. "Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance". In: *Journal of the American Medical Informatics Association* 5.4 (1998), pp. 373–381. DOI: 10.1136/jamia.1998.0050373.

[7] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu. "Recurrent Neural Networks for Multivariate Time Series with Missing Values". In: *arXiv preprint arXiv:1606.01865* (2016).

[8] Y.-T. Cheng, Y.-F. Lin, K.-H. Chiang, and V. S. Tseng. "Mining disease sequential risk patterns from nationwide clinical databases for early assessment of chronic obstructive pulmonary disease". In: *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE. 2016, pp. 324–327.

[9] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun. "GRAM: Graph-based Attention Model for Healthcare Representation Learning". In: *arXiv preprint arXiv:1611.07012* (2016).

[10] E. Choi, M. T. Bahadori, and J. Sun. "Doctor AI: Predicting Clinical Events via Recurrent Neural Networks". In: *arXiv preprint arXiv:1511.05942* (2015).

[11] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart. "RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism". In: *Advances in Neural Information Processing Systems*. 2016, pp. 3504–3512.

[12] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun. "Using recurrent neural network models for early detection of heart failure onset". In: *Journal of the American Medical Informatics Association* (2016), ocw112.

[13] J. Cohen. "A coefficient of agreement for nominal scales". In: *Educational and psychological measurement* 20.1 (1960), pp. 37–46.

[14] J. Cohen. "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit." In: *Psychological bulletin* 70.4 (1968), p. 213.

[15] G. Dong and J. Li. "Efficient mining of emerging patterns: Discovering trends and differences". In: *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. Citeseer. 1999, pp. 43–52.

[16] B. Druss and E. Walker. *Mental disorders and medical comorbidity*.

[17] C. Esteban, O. Staeck, Y. Yang, and V. Tresp. "Predicting Clinical Events by Combining Static and Dynamic Information Using Recurrent Neural Networks". In: *CoRR* abs/1602.02685 (2016). URL: `http://arxiv.org/abs/1602.02685`.

[18] K. M. Flegal, B. K. Kit, and B. I. Graubard. "Body mass index categories in observational studies of weight and risk of death". In: *American journal of epidemiology* 180.3 (2014), pp. 288–296.

[19] J. L. Fleiss, B. Levin, and M. C. Paik. *Statistical methods for rates and proportions*. John Wiley & Sons, 2013.

[20] G. Gariepy, D. Nitka, and N. Schmitz. "The association between obesity and anxiety disorders in the population: a systematic review and meta-analysis". In: *International journal of obesity* 34.3 (2010), p. 407.

[21] L. Geng and H. J. Hamilton. "Interestingness measures for data mining: A survey". In: *ACM Computing Surveys (CSUR)* 38.3 (2006), p. 9.

[22] S. Ghosh, M. Feng, H. Nguyen, and J. Li. "Hypotension Risk Prediction via Sequential Contrast Patterns of ICU Blood Pressure". In: *IEEE Journal of Biomedical and Health Informatics* 20.5 (2016), pp. 1416–1426. ISSN: 2168-2194. DOI: 10.1109/JBHI.2015.2453478.

[23] M. L. Gillison, W. M. Koch, R. B. Capone, M. Spafford, W. H. Westra, L. Wu, M. L. Zahurak, R. W. Daniel, M. Viglione, D. E. Symer, et al. "Evidence for a causal association between human papillomavirus and a subset of head and neck cancers". In: *Journal of the National Cancer Institute* 92.9 (2000), pp. 709–720.

[24] W. R. Gove. "Gender differences in mental and physical illness: The effects of fixed roles and nurturant roles". In: *Social Science & Medicine* 19.2 (1984), pp. 77–84.

[25] K. Gwet et al. "Inter-rater reliability: dependency on trait prevalence and marginal homogeneity". In: *Statistical Methods for Inter-Rater Reliability Assessment Series* 2 (2002), pp. 1–9.

[26] K. L. Gwet. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.

[27] K. A. Hallgren. "Computing inter-rater reliability for observational data: an overview and tutorial". In: *Tutorials in quantitative methods for psychology* 8.1 (2012), p. 23.

[28] W. Hämäläinen. "Kingfisher: an efficient algorithm for searching for both positive and negative dependency rules with statistical significance measures". In: *Knowledge and Information Systems* 32.2 (2011), pp. 383–414. ISSN: 0219-3116.

[29] J. Han, J. Pei, and Y. Yin. "Mining frequent patterns without candidate generation". In: *ACM SIGMOD Record*. Vol. 29. 2. ACM. 2000, pp. 1–12.

[30] D. A. Hanauer and N. Ramakrishnan. "Modeling temporal relationships in large scale clinical associations". In: *Journal of the American Medical Informatics Association* 20.2 (2013), pp. 332–341.

[31] Healthcare Cost and Utilization Project. *Clinical Classifications Software (CCS) for ICD-9-CM*.

[32] D. Heckerman, C. Meek, and G. Cooper. "A Bayesian approach to causal discovery". In: *Innovations in Machine Learning*. Springer, 2006, pp. 1–28.

[33] A. B. Hill. "The environment and disease: association or causation?" In: *Proceedings of the Royal Society of Medicine* 58.5 (1965), pp. 295–300.

[34] K. Järvelin and J. Kekäläinen. "Cumulated Gain-based Evaluation of IR Techniques". In: *ACM Transactions on Information Systems* 20.4 (Oct. 2002), pp. 422–446. ISSN: 1046-8188.

[35] X. Ji, J. Bailey, and G. Dong. "Mining minimal distinguishing subsequence patterns with gap constraints". In: *Knowledge and Information Systems* 11.3 (2007), pp. 259–286.

[36] R. Kavuluru and Y. Lu. "Leveraging output term co-occurrence frequencies and latent associations in predicting medical subject headings". In: *Data & Knowledge Engineering* 94.Part B (2014), pp. 189–201.

[37] R. Kavuluru, A. Rios, and Y. Lu. "An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records". In: *Artificial intelligence in medicine* 65.2 (2015), pp. 155–166.

[38] R. C. Kessler, P. Berglund, W. T. Chiu, O. Demler, S. Heeringa, E. Hiripi, R. Jin, B.-E. Pennell, E. E. Walters, A. Zaslavsky, and H. Zheng. "The US National Comorbidity Survey Replication (NCS-R): design and field procedures". In: *International Journal of Methods in Psychiatric Research* 13.2 (2004), pp. 69–92.

[39] K. Lasser, J. W. Boyd, S. Woolhandler, D. U. Himmelstein, D. McCormick, and D. H. Bor. "Smoking and mental illness: a population-based prevalence study". In: *Jama* 284.20 (2000), pp. 2606–2610.

[40] B. Letham, C. Rudin, and D. Madigan. "Sequential event prediction". In: *Machine learning* 93.2-3 (2013), pp. 357–380.

[41] J. Li, T. D. Le, L. Liu, J. Liu, Z. Jin, B. Sun, and S. Ma. "From observational studies to causal rule mining". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 7.2 (2015), p. 14.

[42] J. Li, T. D. Le, L. Liu, J. Liu, Z. Jin, B. Sun, and S. Ma. "From observational studies to causal rule mining". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 7.2 (2016), p. 14.

[43] L. Li, W.-Y. Cheng, B. S. Glicksberg, O. Gottesman, R. Tamler, R. Chen, E. P. Bottinger, and J. T. Dudley. "Identification of type 2 diabetes subgroups through topological analysis of patient similarity". In: *Science translational medicine* 7.311 (2015), 311ra174–311ra174.

[44] Z. C. Lipton, D. C. Kale, C. Elkan, and R. C. Wetzel. "Learning to Diagnose with LSTM Recurrent Neural Networks". In: *CoRR* abs/1511.03677 (2015). URL: http://arxiv.org/abs/1511.03677.

[45] Y. S. Low, B. Gallego, and N. H. Shah. "Comparing high-dimensional confounder control methods for rapid cohort studies from electronic health records". In: *Journal of comparative effectiveness research* 5.2 (2016), pp. 179–192.

[46] D. D. McAlpine and D. Mechanic. "Utilization of specialty mental health care among persons with severe mental illness: the roles of demographics, need, insurance, and risk." In: *Health services research* 35.1 Pt 2 (2000), p. 277.

[47] M. L. McHugh. "Interrater reliability: the kappa statistic". In: *Biochemia medica* 22.3 (2012), pp. 276–282.

[48] R. A. Miech, A. Caspi, T. E. Moffitt, B. R. E. Wright, and P. A. Silva. "Low socioeconomic status and mental disorders: a longitudinal study of selection and causation during young adulthood". In: *American journal of Sociology* 104.4 (1999), pp. 1096–1131.

[49] S. Moens, E. Aksehirli, and B. Goethals. "Frequent itemset mining for big data". In: *Big Data, 2013 IEEE International Conf. on*. IEEE. 2013, pp. 111–118.

[50] T. H. Moore, S. Zammit, A. Lingford-Hughes, T. R. Barnes, P. B. Jones, M. Burke, and G. Lewis. "Cannabis use and risk of psychotic or affective mental health outcomes: a systematic review". In: *The Lancet* 370.9584 (2007), pp. 319–328.

[51] J. A. Morris and M. J. Gardner. "Statistics in Medicine: Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates". In: *British Medical Journal* 296.6632 (1988), pp. 1313–1316.

[52] M. Mukaka. "A guide to appropriate use of correlation coefficient in medical research". In: *Malawi Medical Journal* 24.3 (2012), pp. 69–71.

[53] B. Nebel and H.-J. Bürckert. "Reasoning about temporal relations: a maximal tractable subclass of Allen's interval algebra". In: *Journal of the ACM (JACM)* 42.1 (1995), pp. 43–66.

[54] W. Opstelten, G. A. Van Essen, F. Schellevis, T. J. Verheij, and K. G. Moons. "Gender as an independent risk factor for herpes zoster: a population-based prospective study". In: *Annals of epidemiology* 16.9 (2006), pp. 692–695.

[55] C. Ordonez, N. Ezquerra, and C. A. Santana. "Constraining and summarizing association rules in medical data". In: *Knowledge and Information Systems* 9.3 (2006), pp. 259–283.

[56] T. Pham, T. Tran, D. Phung, and S. Venkatesh. "Predicting healthcare trajectories from medical records: A deep learning approach". In: *Journal of biomedical informatics* 69 (2017), pp. 218–229.

[57] J. Reps, J. M. Garibaldi, U. Aickelin, D. Soria, J. E. Gibson, and R. B. Hubbard. "Discovering sequential patterns in a UK general practice database". In: *Proceedings of 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics*. IEEE. 2012, pp. 960–963.

[58] S. Robertson. "Understanding inverse document frequency: on theoretical arguments for IDF". In: *Journal of documentation* 60.5 (2004), pp. 503–520.

[59] B. Rosner. *Fundamentals of biostatistics*. Cengage Learning, 2015.

[60] E. Ryu, A. M. Chamberlain, R. S. Pendegraft, T. M. Petterson, W. V. Bobo, and J. Pathak. "Quantifying the impact of chronic conditions on a diagnosis of major depressive disorder in adults: a cohort study using linked electronic medical records". In: *BMC psychiatry* 16.1 (2016), p. 1.

[61] B. Schneider, B. Bartusch, A. Schnabel, and J. Fritze. "Age and gender: confounders for axis I disorders as risk factors for suicide". In: *Psychiatrische Praxis* 32.4 (2005), pp. 185–194.

[62] W. R. Shadish, T. D. Cook, and D. T. Campbell. *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Pub., 2002.

[63] N. H. Shah, P. LePendu, A. Bauer-Mehren, Y. T. Ghebremariam, S. V. Iyer, J. Marcus, K. T. Nead, J. P. Cooke, and N. J. Leeper. "Proton pump inhibitor usage and the risk of myocardial infarction in the general population". In: *PLoS One* 10.6 (2015), e0124653.

[64] I. N. M. Shaharanee, F. Hadzic, and T. S. Dillon. "Interestingness measures for association rules based on statistical validity". In: *Knowledge-Based Systems* 24.3 (2011), pp. 386–392.

[65] G. D. Smith and S. Ebrahim. "Data dredging, bias, or confounding: They can all get you into the BMJ and the Friday papers". In: *BMJ: British Medical Journal* 325.7378 (2002), p. 1437.

[66] C. Spearman. "The proof and measurement of association between two things". In: *The American journal of psychology* 15.1 (1904), pp. 72–101.

[67] P. Spirtes. “Introduction to causal inference”. In: *Journal of Machine Learning Research* 11.May (2010), pp. 1643–1662.

[68] P.-N. Tan, V. Kumar, and J. Srivastava. “Selecting the Right Interestingness Measure for Association Patterns”. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’02. Edmonton, Alberta, Canada: ACM, 2002, pp. 32–41. ISBN: 1-58113-567-X.

[69] B. Taylor, E. Miller, C. Farrington, M.-C. Petropoulos, I. Favot-Mayaud, J. Li, and P. A. Waight. “Autism and measles, mumps, and rubella vaccine: no epidemiological evidence for a causal association”. In: *The Lancet* 353.9169 (1999), pp. 2026–2029.

[70] G. Teodoro, N. Mariano, W. Meira Jr, and R. Ferreira. “Tree projection-based frequent itemset mining on multicore cpus and gpus”. In: *Computer Architecture and High Performance Computing (SBAC-PAD), 2010 22nd International Symposium on*. IEEE. 2010, pp. 47–54.

[71] R. Tibshirani. “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), pp. 267–288.

[72] T. Uno, M. Kiyomi, and H. Arimura. “LCM Ver.3: Collaboration of Array, Bitmap and Prefix Tree for Frequent Itemset Mining”. In: *Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations*. OSDM ’05. Chicago, Illinois: ACM, 2005, pp. 77–86. ISBN: 1-59593-210-0.

[73] D. S. Wald, N. J. Wald, J. K. Morris, and M. Law. “Folic acid, homocysteine, and cardiovascular disease: judging causality in the face of inconclusive trial evidence”. In: *Bmj* 333.7578 (2006), pp. 1114–1117.

[74] G. I. Webb and J. Vreeken. “Efficient discovery of the most interesting associations”. In: *ACM Trans. on Knowledge Discovery from Data* 8.3 (2014), p. 15.

[75] N. Wongpakaran, T. Wongpakaran, D. Wedding, and K. L. Gwet. “A comparison of Cohen’s Kappa and Gwet’s AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples”. In: *BMC medical research methodology* 13.1 (2013), p. 61.

[76] A. Wright, A. McCoy, S. Henkin, M. Flaherty, and D. Sittig. “Validation of an Association Rule Mining-Based Method to Infer Associations Between Medications and Problems”. In: *Applied Clinical Informatics* 4.1 (2013), p. 100.

[77] A. Wright, E. S. Chen, and F. L. Maloney. “An automated technique for identifying associations between medications, laboratory results and problems”. In: *J. of Biomedical Informatics* 43.6 (2010), pp. 891–901.

[78] A. P. Wright, A. T. Wright, A. B. McCoy, and D. F. Sittig. “The use of sequential pattern mining to predict next prescribed medications”. In: *Journal of biomedical informatics* 53 (2015), pp. 73–80.

[79] H. Yang, C. Rudin, and M. Seltzer. “Scalable Bayesian Rule Lists”. In: *arXiv preprint arXiv:1602.08610* (2016).

[80] X. Yin and J. Han. “CPAR: Classification based on Predictive Association Rules.” In: *SIAM International Conf. on Data Mining*. Vol. 3. 2003, pp. 331–335.

[81] M. J. Zaki. “Scalable algorithms for association mining”. In: *Knowledge and Data Engineering, IEEE Transactions on* 12.3 (2000), pp. 372–390.

[82] F. Zhang, Y. Zhang, and J. Bakos. “Gpapriori: Gpu-accelerated frequent itemset mining”. In: *Cluster Computing (CLUSTER), 2011 IEEE International Conference on*. IEEE. 2011, pp. 590–594.

Vita

Name

Orhan Abar

Education

- 2005–2009 B.S. in Computer Engineering Firat University Elazig, Turkey
- 2011–2013 M.S. in Computer Science University of Texas at San Antonio San Antonio, Texas

Experience

- 2018–present, Graduate Research Assistant, University of Kentucky, Lexington, Kentucky

Awards

- 2009-2018, Full-Scholarship from Turkish Government for Graduate Study in the USA

Publications

1. G. L. Heileman, W. G. Thompson-Arjona, H. W. Free and **O. Abar**. Does Curricular Complexity Imply Program Quality? Proceedings of the 2019 American Society for Engineering Education (ASEE) Annual Conference, Tampa, June 16–10, 2019.
2. **Abar, O.**, Charnigo, R. J., Rayapati, A., & Kavuluru, R. On Interestingness Measures for Mining Statistically Significant and Novel Clinical Associations from EMRs. In Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Info. (pp. 587-594). ACM (2016, October).